An Introduction to

ECONOMETRICS

Jaydeb Sarkhel

Santosh Kumar Dutta

*Written according to the new **C.B.C.S.** syllabus of B.A./B.Sc. Honours Course in Economics of different universities in India.*

AN INTRODUCTION TO
ECONOMETRICS

B.A./B.Sc. Economics (Honours)

Jaydeb Sarkhel

Retired Professor, Department of Commerce,
Burdwan University;
Author of 'Microeconomic Theorey', 'Macroeconomic Theory', etc.

Dr. Santosh Kumar Dutta

Associate Professor, Department of Economics,
Bidhannagar College;
Joint Author of 'An Insight into Statistics',
'Microeconomics and Statistics' etc.

Revised and Enlarged
Second Edition
2020

BOOK SYNDICATE (P) LTD.
www.booksyndicate.in

CBCS (UG) Syllabus on Econometrics for different Universities in West Bengal

Paper 1.3 CALCUTTA UNIVERSITY

Economics Core Course X, Core T_{10}
—Introductory Econometrics **(Sem-IV)**, F.M : 100
(Th : 50 + Tutorial : 30 + Internal assessment : 10 + Attendence : 10)

1. Nature and Scope of Econometrics **2 Lecture hours**
1.1 What is Econometrics ?
1.2 Distinction between Economic model and Econometric model
1.3 Concept of Stochstic relation
1.4 Role of random disturbance in econometric model
2. Classical Linear Regression Model (simple linear regression and multiple linear regression) : Part I **18 Lecture hours**
2.1 The classical assumptions
2.2 Concepts of population regression function and sample regression function
2.3 Estimation of model by the method of ordinary least squares
3. Classical linear regression model (simple linear regression and multiple linear regression) : Part 2 **15 Lecture hours**
3.1 Properties of Least Squares Estimators (BLUE)-Gauss-Markov theorem
3.2 Qualitative (dummy) independent variables (only interpretation of the model)
3.3 Forecasting (only for two variable model) : Expost forecast and Exante forecast
4. Statistical inference in Linear regression model **20 Lecture hours**
4.1 Sampling distribution of regression estimates : Standard normal, Chi-square, *t*, F
4.2 Confidence intervals
4.3 Concepts of Type I and Type II errors
4.4 Testing of hypothesis about β and σ^2 given and with unknown σ^2 (Standard normal and '*t*' statistics)
4.5 Testing hypothesis involving several parameters : the F test
4.6 Goodness of fit (in terms of R^2, adjusted R^2 and F statistic)
5. Violations of Classical Assumptions **10 Lecture hours**
5.1 Multicollinearity–Consequences, Detection and Remedies
5.2 Heteroscedasticity–Consequences, Detection and Remedies
5.3 Autocorrelation–Consequences, Detection and Remedies
6. Specification Analysis **10 Lecture hours**
6.1 Omission of a relevant variable
6.2 Inclusion of an irrelevant variable
6.3 Tests of specification errors
6.4 Testing for linearity and normality assumptions.

CONTENTS

1

Definition, Scope and Goals of Econometrics

1.1. Definition and Scope of Econometrics

Literally speaking, the word 'econometrics' means "measurement in economics". Econometrics may be considered as the integration of economics, mathematics and statistics for the purpose of providing numerical values for the parameters of economic relationships and verifying economic theories. It is a special type of economic analysis in which the general economic theory formulated in mathematical terms is combined with empirical measurement of economic phenomena. We start from general economic theory, that is, from the relationships of economic variables as suggested by economic theory and express them in mathematical terms. This is called building of an economic model. Next we use statistical methods in order to obtain numerical estimates of the coefficients of the economic relationships. These statistical methods are called ***econometric methods.***

Although measurement is an important part of econometrics, the scope of econometrics is much broader, as can be seen from the following quotations :

"Econometrics, the result of a certain outlook on the role of economics, consists of the application of mathematical statistics to economic data to lend empirical support to the models constructed by mathematical economics and to obtain numerical results"[1].

"Econometrics may be defined as the quantitative analysis of actual economic phenomena based on the concurrent development of theory and observation, related by appropriate methods of inference."[2]

"Econometrics may be defined as the social science in which the tools of economic theory, mathematics and statistical inference are applied to the analysis of economic phenomena."[3]

"Econometrics is concerned with the empirical determination of economic laws."[4]

"The art of the econometrician consists in finding the set of assumptions that are both sufficiently specific and sufficiently realistic to allow him to take the best possible advantage of the data available to him."[5]

"Econometricians ... are of positive help in trying to dispel the poor public image of economics (quantitative or otherwise) as a subject in which empty boxes are opened

[1] Gerhard Tintner, *Methodology of Mathematical Economics and Econometrics*, The University of Chicago Press, Chicago, 1968, p. 74.

[2] P. A. Samuelson, T. C. Koopmans, and J. R. N. Stone, "Report of the Evaluative Committe for Econometrica," *Econometrica*, vol. 22, no. 2, April 1954, pp. 141-146.

[3] Arthur S. Goldberger, *Econometric Theory*, John Wiley & Sons, New York, 1964, p. 1.

[4] H. Theil, *Principles of Econometrics*, John Wiley & Sons. New York, 1971, P. 1.

[5] E. Malinvaud, *Statistical Methods of Econometrics*, Rand Mc Nally, Chicago, 1966, P. 514.

by assuming the existence of can-openers to reveal contents which any ten economists will interpret in 11 ways."[6]

"The method of econometric research aims, essentially, at a conjunction of economic theory and actual measurements, using the theory and technique of statistical inference as a bridge pier."[7]

All these definitions suggest that econometrics is an amalgam of economic theory, mathematical economics, economic statistics and mathematical statistics.

Economic theory postulates an exact relationship between economic variables but actually an economic relationship always contains a random element. Economic theory ignores it but econometrics does not, because the econometric methods can deal with these random components. For example, in the Keynesian macroeconomic theory we find an exact relationship between consumption expenditure (C) and income (Y). Keynesian consumption function is given by $C = a + bY$ where $a > 0$ is called the autonomous part of consumption expenditure and $b = \frac{dC}{dY}$ is called the marginal propensity to consume (MPC) [$0 < b < 1$ by assumption]. This is an exact relationship because C is completely determined by Y. So, in this model the effects of other variables like price, wealth, income distribution etc., are ignored. But in econometrics the influence of other factors is considered by introducing a random variable in the model and that random variable is generally denoted by 'u' called error term. So, the consumption function considered in econometrics is $C = a + bY + u$. Now econometric methods estimate the parameters 'a' and 'b' and while estimating 'a' and 'b' the choice of the econometric method depends on the behaviour of the distribution of the random variable 'u'.

There are three main sources of the error term 'u' in the functional relation. These are :

(i) unpredictable element of randomness in human response,

(ii) effect of a large number of variables that have been omitted from the functional relation.

and (iii) measurement error. (For details see Section 2.2.1)

1.2. Relationship between Econometrics and Economic Theory

Economic theory makes statements or hypotheses that are mostly qualitative in nature.

Econometrics presupposes the existence of a body of economic theory. Economic theory should come first because it states the hypothesis about economic behaviour which should be tested with the econometric methods.

For example, we consider the consumption income relationship of the form

$$C = a + bY + u.$$

Economic theory suggests that consumption is a function of income and with the information of economic theory we know that $MPC = \frac{dC}{dY} = b$ lies between 0 and 1 i.e., $0 < b < 1$.

[6] Adrian C. Darnell and J. Lynne Evans, *The Limits of Econometrics*, Edward Elgar Publishing, Hants, England, 1990, P. 54.

[7] T. Haavelmo, "The Probability Approach in Econometrics", Supplement to *Econometrica*, vol. 12, 1944, Preface P. iii.

The proposition suggested by economic theory is to be tested now, applying econometric methods. If we find that the theory is consistent with the empirical results we accept the theory but if we find that it is not consistent with the empirical results, then we have either to reject the theory or to modify the theory. If we like to modify the theory then we should not reject the theory, rather we should incorporate some other variables and parameters to make the theory more meaningful (and close to reality).

For example, the simple consumption income relation, $C = a + bY + u$ can be modified in the form

$C = a + bY + cP + dW + u$ where two new variables P (price level) and W (wealth) have been taken into account in the functional relation. The signs of the parameters (a, b, c, $d > 0$) and the corresponding response coefficients can also be tested empirically.

1.2.1. Difference between Economic Model and Econometric Model

A model is a simplified representation of a real world process. In practice, in any economic model (say consumption function or demand function), we can include all the relevant variables that we think are relevant for our purpose and dump the rest of the variables in a basket called "disturbance". This brings us to the distinction between an economic model and an econometric model.

An economic model is a set of assumptions that approximately describe the behaviour of an economy. An econometric model, on the other hand, consists of the following :

(i) A set of behavioural equations derived from the economic model.

(ii) A statement of whether there are errors of observation in the observed variables.

(iii) A specification of the probability distribution of the "disturbances"(and errors of measurement).

For example, we may consider a simple demand model of economics. Then econometric model will usually consist of :

(a) The behavioural equation : $q = \alpha + \beta p + u$ where q = quantity demanded, p = price, α and β are two parameters and u = random disturbance term.

(b) A specification of probability distribution of u, where values of u are independently and normally distributed with mean $E(u) = 0$ and variance $(u) = \sigma_u^2$. With these specifications we can test empirically the law of demand or the hypothesis that $\beta < 0$.

We may also use the estimated demand function for prediction and policy purposes.

1.3. Econometrics and Mathematical Economics

Mathematical economics states economic theory in terms of mathematical symbols. There is no essential difference between mathematical economics and economic theory. Both state the same relationships, but while economic theory uses verbal exposition, mathematical economics employs mathematical symbolism. Both express the various economic relationships in an exact form. Neither economic theory nor mathematical economics allows for random elements which might affect the relationship and make it *stochastic*. Furthermore, they do not provide numerical values for the coefficients of the relationships. Relations in economic theory or in mathematical economics are of *non-stochastic* form. It is in this regard that econometrics differs from mathematical economics.

Although econometrics presupposes the expression of economic relationships in mathematical form, like mathematical economics it does not assume that economic relationships are exact. On the contrary, econometrics assumes that relationships are not exact. Econometric methods are designed to take into account random disturbances which create deviations from the exact behavioural patterns suggested by economic theory and mathematical economics. Furthermore, econometric methods provide numerical values of the coefficients of economic phenomena. Thus by combining mathematical formulations of theory with empirical data, econometrics enables us to pass from the abstract theoretical scheme to numerical results in concrete cases.

1.4. Econometrics and Statistics

Econometrics differs both from mathematical statistics and economic statistics. An economic statistician gathers empirical data, records them, tabulates them or charts them, and then attempts to describe the pattern in their development over time and perhaps detect some relationship between various economic magnitudes. Thus economic statistics is mainly a descriptive aspect of economic theory. It does not provide explanations of the development of the various variables and it does not provide measurement of the parameters of economic relationships.

Economic statistics differs from mathematical (or inferential) statistics. Mathematical statistics is based upon the theory of probability and deals with the methods of measurement which are developed on the basis of controlled or carefully planned experiments. These statistical methods cannot be applied in economic relationships because such experiments cannot be designed (except in a very few cases, e.g., agricultural experiments or industrial experimentation) for economic phenomena.

Econometrics uses statistical methods after adopting them to the problems of economic life. These adopted statistical methods are called econometric methods. In particular, econometric methods are so adjusted that they become appropriate for the measurement of economic relationships which are stochastic, that is, they include random elements. The adjustment consists primarily in specifying the stochastic (random) elements that are supposed to operate in the real world and enter into the determination of the observed data, so that the latter can be interpreted as a (random) sample to which the methods of statistics can be applied.

1.5. Goals of Econometrics

Econometrics helps us to achieve the following three main goals :

(i) **Analysis** : This means testing of economic theory. There are alternative theories to explain the functioning of the economic system. Econometrics examines the explanatory power of the system.

(ii) **Policy making** : The numerical estimates of the coefficients of the economic relationships help the policy-maker to define the apporpriate policies. For example, the numerical estimate of price elasticities of demand for a product will help the policy maker to know how much additional revenue is expected to be obtained if sales tax is imposed on that commodity. Alternatively, numerical estimates of price elasticities of exports and imports will help us to know how far the devaluation as a policy will be effective in solving the balance of payments deficit problem.

(iii) **Forecasting** : The numerical estimates of the coefficients are used in order to forecast the future value of the economic variables. Without forecasting the planner cannot adopt appropriate policies. Of course, these goals are not mutually exclusive.

Successful econometric applications should really include some combination of all those aims.

1.6. Division of Econometrics

Econometrics may be divided into two branches : theoretical econometrics and applied econometrics.

Theoretical econometrics includes the development of appropriate methods for the measurement of economic relationships. Econometric techniques are basically statistical techniques which have been adopted to the particular characteristics of economic relationships.

Econometric methods may be classified into two groups : (i) Single-equation techniques, which are methods that are applied to one relationship at a time, and (ii) Simultaneous equation techniques, which are methods applied to all the relationships of a model simultaneously.

Applied econometrics includes the applications of econometric methods to specific branches of economic theory. It examines the problems encountered and the findings of applied reasearch in the fields of demand, supply, production, investment, consumption, and other sectors of economic theory. Applied econometrics involves the application of the tools of theoretical econometrics for the analysis of economic phenomena and forecasting economic behaviour.

1.7. Methodology/Stages of Econometric Research

Applied econometric research is concerned with the measurement of the parameters of economic relationships and with the prediction of the values of economic variables.

The relationships of economic theory which can be measured with one or another econometric techniques are causal, that is, they are relationships in which some variables are postulated as causes of the variation of other variables. In this sense definitional equations do not require any measurement. For example, the equation $Y = C + I$ is the mathematical expression of the definition of national income (in a closed economy, with no government activity) of economic theory. It does not explain the determination of the level of income or the causes of its variations.

There are four stages in any econometric research.

Stage A : Specification of the model :

It means expressing the relationships between the variables in mathematical form. This stage is also called formulation of the maintained hypothesis. It involves the determination of

(i) dependent and the explanatory variables to be included in the model.

(ii) the theoretical expectations about the sign, size of the parameters of the function.

(iii) the mathematical form of the model.

For example, consider a production function of the following type $Y = f(K, L)$ where K and L are the two factors of production.

[K = Capital, L = Labour and Y is the level of output.] This function can also be written in the Cobb-Douglas form i.e., $Y = K^{\alpha} L^{\beta}$ or, $\log Y = \alpha \log K + \beta \log L$. This is the mathematical form of log linear function. Here some theoretical restrictions must be imposed : $0 < \alpha, \beta < 1$,

$\alpha + \beta > 1$ if there are increasing returns to scale

$\alpha + \beta < 1$ if there are decreasing returns to scale

$\alpha + \beta = 1$ if there are constant returns to scale.

α = Elasticity of output with respect to capital

β = Elasticity of output with respect to labour.

Stage B : Estimation of the model :

After the model has been specified (formulated) the econometrician must proceed with its estimation. The estimation of the model is a purely technical stage.

This stage includes the following steps :

(i) Collection of data on different variables included in the model.

(ii) Examination of the identification conditions of the function in which we are interested.

(iii) Examination of aggregation problems involved in the variables of the function.

(iv) Examination of the degree of correlations among the explanatory variables.

Let us consider the relationship $C = \beta_0 + \beta_1 P + \beta_2 Y + \beta_3 W + u$ where C = consumption expenditure (dependent variable) and explanatory variables are : Y = income, W = wealth, P = price level. Now we have to find out whether there exists any correlations among the explanatory variables (i.e. correlation between Y and P or Y and W or P and W i.e., the problem of multicollinearity.)

(v) Choice of the appropriate econometric technique for the estimation of the function and critical examination of the assumptions of the chosen technique and of their economic implications for the estimates of the coefficients.

Stage C : Evaluation of estimates

After the estimation of the model the econometrician must proceed with the evaluation of the results of the calculations, that is with the determination of the reliability of these results. The evaluation consists of deciding whether the estimates of the parameters are theoretically meaningful and statistically significant.

For this purpose we may use various criteria which may be classified into three main groups :

(i) **Economic criteria** : These are determined by the principles of economic theory and refer to the sign and the size of the parameters of economic relationships. For example, the Keynesian liquidity preference function may be expressed in the mathematical form

$$M = \beta_0 + \beta_1 Y + \beta_2 r + u$$

where M = demand for money (dependent variable), Y = income, r = rate of interest, u = error term, β_0, β_1, β_2 are the parameters whose values and signs are to be determined on the basis of observed data. On the basis of the existing theory the signs of the parameters would be : $\beta_0 > 0$, $\beta_1 > 0$, $\beta_2 < 0$.

(ii) **Statistical criteria (First order tests)** : These are determined by statistical theory and aim at the evaluation of the statistical reliability of the estimates of the parameters of the model. The most widely used statistical criteria are the correlation coefficient and the standard error of the estimates.

(iii) **Econometric criteria (Second order tests)** : These are set by the theory of econometrics and aim at the investigation of whether the assumptions of the econometric method employed are satisfied or not in any particular case. The econometric criteria serve as second order tests (as tests are the statistical tests) : in other words they determine the reliability of the statistical criteria, and in particular of the standard errors of the parameter estimates. They help us to establish whether the estimates have the desirable properties of unbiasedness, consistency etc.

Stage D : Evaluation of the forecasting power of the estimated model

The objective of any econometric research is to obtain good numerical estimates of the coefficients of economic relationships and to use them for the prediction of the values of economic variables. Estimates of the parameters are useful because these help the policy makers in adopting the policies. A model after the estimation of the parameters can be used in forecasting the values of the economic variables. So the econometricians must test the forecasting power of the model.

1.8. Desirable Properties of an Econometric Model

The 'goodness' of an econometric model is judged customarily according to the following desirable properties :

(i) **Theoretical plausibility :** The model should be compatible with the postulates of economic theory. It must describe adequately the economic phenomena to which it relates.

(ii) **Explanatory ability :** The model should be able to explain the observations of the actual world. It must be consistent with the observed behaviour of the economic variables whose relationship it determines.

(iii) **Accuracy of the estimates of the parameters :** The estimates of the coefficients should be accurate in the sense that they should approximate as best as possible the true parameters of the structural model.

(iv) **Forecasting ability :** The model should produce satisfactory predictions of future values of the dependent variables.

(v) **Simplicity :** The model should represent the economic relationships with maximum simplicity.

1.9. Nature and Sources of Data for Economic Analysis

The success of any econometric analysis depends on the availability of the appropriate data. Three types of data are generally available for empirical analysis : *time series data, cross-section data* and *pooled data/panel data.*

Time Series Data

A time series is a set of observations on the values that a variable takes at different times. Such data may be collected at regular time intervals, such as *daily* (e.g., stock prices, weather reports), *weekly* (e.g. money supply figures), *monthly* (e.g., unemployment rate, Consumer Price Index (CPI), *quarterly* (e.g., GDP), *anually* (e.g. goverment budget), *quinquennially,* that is, every 5 years (e.g., the census of manufactures) or *decennially,* that is every 10 years (e.g., the census of population).

Cross-Section Data

Cross-section data are data on one or more variables collected at the same point of time, such as the census of population conducted by the Government of India every 10 years, the Survey of household consumer expenditure in India conducted by National Sample Survey Organization (NSSO), the opinion polls by the Times of India, NDTV, CNN-IBN and many other organizations. An individual researcher or a group may also collect cross-section data directly from the field of enquiry.

Conventionally the letter Y denotes the dependent variable and X's ($X_1, X_2, ..., X_K$) denote the explanatory/independent variables, X_K being the *Kth* explanatory variable. The subscript i or t denote *ith* or the *t th* observation or value. X_{Ki} (or X_{Kt}) will denote the *ith* (or *t th*) observation on variable X_K. Here N (or T) will denote the total number of observations or values in the population and n (or t) will denote the total number of observations in a sample. Normally, the subscript i will be used for cross-section data (i.e., data collected at one point of time) and the subscript t will be used for time series data (i.e., data collected on different periods of time). For instance, consider the Keynesian consumption function of the form $C = a + bY$ where C = consumption expenditure, Y = income and a and b are two constants, a = autonomous part of consumption expenditure, b = marginal propensity to consume. According to the existing theory $a > 0$, $0 < b < 1$. If we like to test this relation with the help of time series data, then we will write the regression equation in the form $C_t = a + bY_t + u_t$ (where $t = 1, 2, ..., t$ (say)) where u is the random disturbance term. On the other hand, we can write the regression equation in the form $C_i = a + bY_i + u_i$, $i = 1, 2, ..., n$ (say), when we verify the relation with the help of cross-section data.

Pooled Data

Pooled or combined data are elements of both time series and cross-section data. Generally speaking, pooled data is a combination of data (i.e., sales, advertisement, earnings etc) of say 20 firms over a given period of time say, a year or two. These combined data of 20 firms in 2 years making 40 observations make it a pooled data, that is, pooling 20 firms' data in 2 years together. So, it is a combination of cross section data and time series data.

Panel, Longitudinal, or Micropanel Data

This is a special type of pooled data in which the same cross-sectional unit (say, a family or a firm) is surveyed over time.

For example, the Nigerian population commission surveys each house every 10 years to determine the changes that may have occurred within these years. By surveying or interviewing the same households or firms to find out their population or financial conditions periodically (10 years interval), panel data can help to provide useful information on the changes that may have occurred in these households. It's more detailed than just the pooled data in a short period of time.

Thus, there is a basic difference between pooled data and panel data. It should be noted that pooled time-series, cross-section data are data with relatively few cross-sections (say few firms under study), where variables are held in cross-section specific individual series (i.e. sales, advertisement, earnings, etc.), while panel data correspond to data with large number of cross-sections, with variables held in single series in stacked form.

The Sources of Data

The data used in empirical analysis may be collected by a government agency (e.g., the Central Statistical Organization), an international agency (e.g., the International Monetary Fund (IMF) or the World Bank), a private organization (e.g., the Centre for Monitoring Indian Economy) or an individual. There exist a lot of agencies collecting data for one purpose or another. Now a days the Internet has revolutionized data gathering. Most of the data can be downloaded from different websites either free of cost or with minimum cost.

The Accuracy of Data

Although plenty of data are available for economic research, the quality of data is often not that good.

There are several reasons for that :

(i) Most of the social science data are non experimental in nature. Therefore there is the possibility of observational errors.

(ii) Even in experimentally collected data, errors of measurement arise from approximations and rounding offs.

(iii) In questionnaire type of surveys, the problem of non-response may lead to bias in results.

(iv) The sampling methods used in obtaining data may vary so widely that it is often difficult to compare the results obtained from the various samples.

(v) Economic data are generally available at a highly aggregate level. Such highly aggregated data may not be helpful for individualistic study.

Because of all of these and many other problems, the researchers should always keep in mind that the results of research are only as good as the quality of the data. Therefore, if in given situations researchers find that the results of the research are "unsatisfactory" the cause may not be that they used the wrong model but due to the poor quality of data.

1.10. A Note on the Measurement Scales of Variables

The variables that we generally use, can be measured in four types of scales : *ratio scale, interval scale, ordinal scale* and *nominal scale*. We can briefly describe them as follows :

Ratio Scale : For a variable X, taking two values say X_1 and X_2, the ratio X_1/X_2 and the distance $(X_2 - X_1)$ are meaningful quantities. Also, there is a natural ordering (ascending or descending) of the values along the scale (say $X_2 \geq X_1$ or $X_2 \leq X_1$). Most economic variables belong to this category. Personal income, measured in rupees is a ratio variable, someone earning ₹ 50,000 is making twice as much as another person earning ₹ 25000.

Interval scale : The interval scale satisfies the last two properties (stated in ratio scale) but not the first.

For example, the distance between two time periods, say (2018-2001) is meaningful but not the ratio of two time perids $\left(\frac{2018}{2001}\right)$.

Ordinal Scale : A variable belongs to this category only if it satisfies the third property of the ratio scale (i.e., natural ordering). Examples are grading systems (A, B, C, grades) or income class (upper, middle, lower). For these variables the ordering exists. But the distances between the categores cannot be quantified.

Nominal Scale : Variables in this category have more of the features of the ratio scale variables. Variables such as gender (male, female) and marital status (married, unmarried, divorced, separated) simply denote categories.

EXERCISE

1. What is Econometrics and what are its components ? Describe the function of each component. Give examples in support of your answer.

2. Name and describe three relationships studied in Economic theory which can be estimated as subject-matter of Econometrics. What are the parameters of these relationships ?

3. How would you define Econometrics ? How does it differ from Mathematical Economics and Statistics ? Describe the main steps involved in any econometric research by taking an example from economic theory.

4. Considering the following relations, how would you explain that economic theory postulates exact relationships between economic variables. How can these be transformed into econometric relations ?

   (i) Demand function : $D = \alpha + \beta_1 P + \beta_2 Y$ where D = quantity demanded, P = price and Y = income.

   (ii) Supply function : $S = \alpha + \beta P$ where S = quantity supplied, P = price.

   (iii) Consumption function : $C = \alpha + \beta Y_d$ where C = consumption expenditure and Y_d = disposable income.

   (iv) Cost function $C = \alpha + \beta X$ where C = total cost and X = total output.

   (v) Production function : $Q = AL^{\alpha} K^{\beta}$ where Q = level of output, L = labour input, K = capital input, A = constant technical parameter, α, β are the two elasticity coefficients.

   (a) What is the economic meaning of the coefficients involved in all the above relations?

   (b) What would you expect about the sign and size of the coefficients to be in each of the above relationships ?

5. Enumerate the relation between econometrics and economic theory.

6. What is Econometrics ? What are the different goals of econometrics ?

7. Distinguish between theoretical econometrics and applied econometrics.

8. Explain briefly the different stages of any econometric research.

9. What is an econometric model ? Illustrate any one of such models.

10. What are the desirable properties of an econometric model ?

11. What are the different types of data available for empirical analysis ?

12. Distinguish between time series data and cross-section data.

13. What are pooled data ? What are panel data ? Distinguish between pooled data and panel data.

14. What are the different sources of data used in empirical analysis ?

15. What do you mean by accuracy of data ? What are the different reasons for distortion of accuracy of data collected and published by different organizations ?

16. Give a brief outline on measurement scales of variables.

2

The Simple Linear Regression Model

2.1. Introduction

Most of economics is concerned with relations among variables. These relations when phrased in mathematical terms can predict the effect of one variable on another. For example, assuming that income, prices of other commodities and all other determinants of demand are constants, we can express the quantity demanded (q) of any commodity as a function of the price (p) of that commodity only. This may be put in the form $q = f(p)$. Similarly we are familiar with other functions (with different assumptions) such as consumption function : $C = f(Y)$, supply function : $S = f(p)$, cost function $C = f(q)$, production function : $Q = f(x_1, x_2)$ where x_1 and x_2 are amounts of different inputs, etc.

These functional relationships define the dependence of the dependent variable upon the independent variable (s) in the specific form. The functional relation may be linear, quadratic, logarithmic, exponential or hyperbolic.

A relation between two variables X and Y expressed as $Y = f(X)$ is said to be deterministic or non-stochastic (non-random) if for each value of the independent variable (X) there is one and only one corresponding value of the dependent variable (Y). On the other hand, a relation between X and Y is said to be stochastic if for a particular value of X there is a whole probability distribution of values of Y. In such a case for any given value of X, the dependent variable Y assumes some specific value only with some probability.

For example, a linear demand function (in deterministic form) can be written as : $q = f(p) = \alpha + \beta p$ ($\alpha > 0$, $\beta < 0$) and in particular $q = 100 - 5p$. When $p = 10$, $q = 50$, when $p = 15$, $q = 25$ etc.

But such an exact and deterministic relation between p and q is never true in the real world.

The deterministic behaviour of the above relationship breaks down when the *ceteris paribus* (other things remaining the same) condition is relaxed.

We, therefore, rewrite the demand equation as : $q = \alpha + \beta p + u$ or in particular, $q = 100 - 5p + u$ where 'u' is commonly known as *random disturbance,* since it disturbs an otherwise deterministic relation.

2.1.1. Concepts of Population Regression Function and Sample Regression Function

Sampling denotes the selection of a part of the aggregate statistical material with a view to obtaining information about the whole. This aggregate or totality of statistical information on a particular character of all the members covered by an investigation is called population or universe. When the population size is very large, it may not be possible to take a complete enumeration of the population. Then we select a small part of the population called sample and examining this small part we can infer about the nature of the whole population. The basic objective of sampling is to make inference about the population by examining a small part of it.

In reality we may be interested to find out the relation between two or more variables simultaneously. In the case of simple (linear) regression model we assume only one explanatory variable but in the case of multiple regression model we assume more than one explanatory variables. The first case is known as the bivariate analysis while the second case is known as the multivariate analysis. Initially we will concentrate on bivariate analysis (study the relation between two variables *X* and *Y* only, where *Y* = dependent variable, *X* =independent/explanatory variable).

We know that regression analysis is largely concerned with estimating and/or predicting the population paramater, say mean value of the dependent variable (*Y*) on the basis of the known or fixed values of the explanatory variable(s). To understand the fact we consider a total population of 60 families in a hypothetical community and their monthly income (*X*) and monthly consumption expenditure (*Y*), both in rupees. These 60 families are divided into 10 income groups and the monthly expenditures of each family in the various groups are shown in the following table (Table 2.1)

Table 2.1 : Joint distribution of monthly income (*X* in ₹) and monthly consumption expenditure (*Y* in ₹) of 60 families in a hypothetical community

X→ Y ↓	8000	10000	12000	14000	16000	18000	20000	22000	24000	26000
	5500	6500	7900	8000	10200	11000	12000	13500	13700	15000
	6000	7000	8400	9300	10700	11500	13600	13700	14500	15200
	6500	7400	9000	9500	11000	12000	14000	14000	15500	17500
	7000	8000	9400	10300	11600	13000	14400	15200	16500	17800
	7500	8500	9800	10800	11800	13500	14500	15700	17500	18000
	—	8800	—	11300	12500	14000	—	16000	18900	18500
	—	—	—	11500	—	—	—	16200	—	19100
Total	32500	46200	44500	70700	67800	75000	68500	104300	96600	121100
Conditional means of Y, E (*Y*/*X*)	6500	7700	8900	10100	11300	12500	13700	14900	16100	17300

Here we have 10 fixed values of *X* and the corresponding *Y* values against each of the *X* values and hence we have 10 subpopulations of *Y*. From Table 2.1 we see that there is considerable variation in monthly consumption expenditure in each income group but the general picture is that despite the variability of monthly consumption expenditure within each income bracket, on an average, monthly consumption expenditure increases as income increases. To understand it clearly we have given the mean, or average montly consumption expenditure corresponding to each of the 10 levels of income. Thus, corresponding to the monthly income level of ₹8000, the mean consumption expenditure is ₹6500 and so on. In total we have 10 mean values for 10 sub-populations of *Y* and these mean values are called *conditional expected values* as they depend upon the given values of the (conditioning) variable *X*.

Symbolically, we denote them as $E(Y/X)$ which simply means the expected value of Y given the value of X. It should be noted that these expected values are called conditional expected values. In order to calculate conditional expected values of Y we have to construct conditional probability distribution of Y, $P(Y/X_i)$, shown in Table 2.2.

Table 2.2 : Conditional Probabilities $P(Y/X_i)$ for the data of Table 2.1

X→ $P(Y/X_i)$ ↓	8000	10000	12000	14000	16000	18000	20000	22000	24000	26000
Conditional probabilities $P(Y/X_i)$	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	—	$\frac{1}{6}$	—	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{6}$	—	$\frac{1}{7}$	$\frac{1}{6}$	$\frac{1}{7}$
	—	—	—	$\frac{1}{7}$	—	—	—	$\frac{1}{7}$	—	$\frac{1}{7}$
Conditional means of Y	6500	7700	8900	10100	11300	12500	13700	14900	16100	17300

For the income group of ₹8000 the expected monthly expenditure is obtained as :

$$₹5500 \times \frac{1}{5} + ₹6000 \times \frac{1}{5} + ₹6500 \times \frac{1}{5} + ₹7000 \times \frac{1}{5} + ₹7500 \times \frac{1}{5} = ₹6500.$$

The expected monthly expenditures for other income groups are also obtained in this way.

It is important to distinguish these conditional expected values from the *unconditional expected value* of monthly consumption expenditure, $E(Y)$. If we add the monthly consumption expenditures for all the 60 families in the population and divide this number by 60, we get the value ₹12120 (₹727200/60) which is the unconditional mean or expected value of Y, $E(Y)$.

Thus the expected monthly consumption expenditure of a family would be ₹12120 (the unconditional mean). But if we like to know the expected value of monthly consumption expenditure of a family whose monthly income is say ₹12000, then we get a value of ₹8900 (the conditional mean).

Graphically, if we join these conditional mean values, we obtain the **population regression line (PRL)** or **population regression curve** or simply it is the *regression of Y on X*.

The polulation regression curve is simply the locus of the conditional means of the dependent variable for the fixed values of the explanatory variable(s). More specifically, it is the curve connecting the means of the sub-populations of Y corresponding to the given values of the regressor X. This is shown in Figure 2.1.

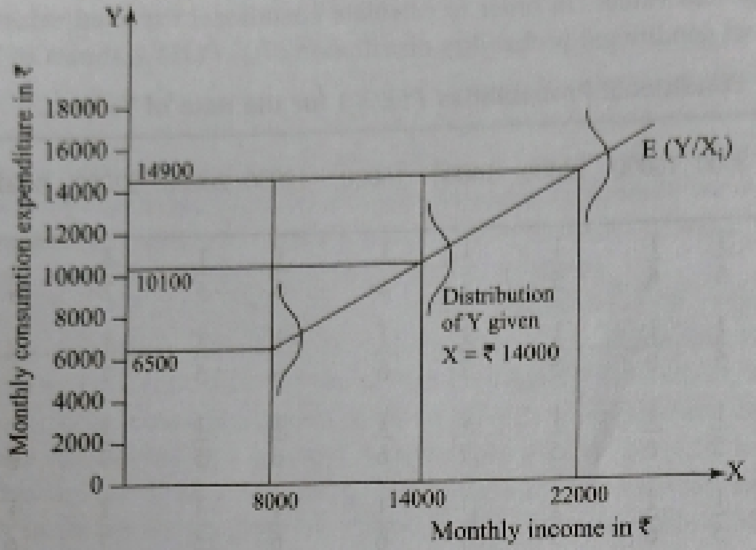


Fig. 2.1. Population regression line (data of Table 2.1)

This figure shows that for each X (i.e., income level) there is a population of Y values (monthly consumption expenditure) that are spread around the (conditional) mean of those Y values. For simplicity we have assumed that these Y values are distributed symmetrically around their respective (conditional) mean values and the regression line (or curve) passes through these (conditional) mean values.

2.1.2. Population Regression Function (PRF)

From the above explanation and Figure 2.1 it is clear that each conditional mean $E(Y/X_i)$ is a function of X_i, where X_i is a given value of X. Symbilically, $E(Y/X_i) = f(X_i)$ where $f(X_i)$ denotes some function of the explanatory variable X. It is a linear function in X_i. This function is also known as the **conditional expectation function *(CEF)* or Population regression function *(PRF)*** or **population regression (*PR*)**. It states merely that the expected value of the distribution of Y given X_i is functionally related to X_i. In simple terms, it tells how the mean or average response of Y varies with X. However, the functional form of the *PRF* is an empirical question. For example, to verify the consumption income relation we generally assume a linear relation. We may assume that the *PRF*, $E(Y/X_i)$ is a linear function of X_i, say, of the type

$E(Y/X_i) = \alpha + \beta X_i$ where α and β are unknown but fixed parameters known as the regression coefficients ; α and β are also known as **intercept** and **slope** coefficients respectively. In regression analysis our interest is in estimating the *PRF* and the unknown values of α and β on the basis of observations on Y and X.

From our earlier example stated in Table 2.1 we see that, given the income level of X_i (say), an individual family's consumption expenditure is clustered around the average consumption of all families at X_i, i.e., around its conditional expectation. Therefore, we can express the deviation of an individual Y_i around its expected value as follows :

$$u_i = Y_i - E(Y/X_i) \text{ or, } Y_i = E(Y/X_i) + u_i \text{ or, } Y_i = \alpha + \beta X_i + u_i$$

where the deviation u_i is an unobservable random variable taking positive or negative values. Technically, u_i is konwn as the **Stochastic disturbance term** or **Stochastic error term.**

2.1.3. The Sample Regression Function (SRF)

In discussing population regression we have deliberately avoided sampling considerations (Data of Table 2.1 represent population, not sample). In most of the cases we estimate the PRF on the basis of the sample information taken from the population. For example, we assume that the population of Table 2.1 is not known to us and the only information we have is a randomly selected sample of Y values for the fixed X's as shown in Table 2.3. Unlike Table 2.1, we have one Y corresponding to the given X's ; each Y (given X_i) in Table 2.3 is chosen randomly for similar Y's corresponding to the same X_i from the population of Table 2.1.

Table 2.3

A random sample from the population of Table 2.1

Y	X
7000	8000
6500	10000
9000	12000
9500	14000
11000	16000
11500	18000
12000	20000
14000	22000
15500	24000
15000	26000

Table 2.4

A random sample from the population of Table 2.1

Y	X
5500	8000
8800	10000
9000	12000
8000	14000
11800	16000
12000	18000
14500	20000
13500	22000
14500	24000
17500	26000

Now from the sample of Table 2.3 we can predict or forecast the average monthly consumption expenditure Y in the population as a whole corresponding to the chosen X's.

However, we may not be able to estimate the PRF "accurately" because of sampling fluctuations. To see this we have drawn another random sample from the same population (Table 2.1) and shown in Table 2.4.

Now plotting the data of Tables 2.3 and 2.4 we obtain the scatter diagram, shown in Figure 2.2.

In the scatter diagram two sample regression lines are drawn so as to 'fit' the scatters reasonably well ; SRF_1 is based on the first sample and SRF_2 is based on the second sample. The regression lines in Figure 2.2 are known as the sample regression lines. Supposedly they represent the population regression line, but due to sampling fluctuations they are at best an approximation of the true *PR*. In general, we may get N different *SRFS* for N different samples and these *SRFS* are not likely the same.

Like the *PRF* (derived from population regression line) we can develop the concept of the **sample regression function *(SRF)*** to represent the sample regression line.

Since from the population regression function we know that the function is of the form, $Y_i = E(Y/X_i) = \alpha + \beta X_i$, the sample counterpart of this equation may be written as

$\hat{Y}_i = \hat{\alpha} + \hat{\beta}\ X_i$ where

$\hat{Y}_i$ = estimator of $E(Y/X_i)$, $\hat{\alpha}$ = estimator of α and $\hat{\beta}$ = estimator of β.

It should be noted that an **estimator**, also known as a (sample) *statistic*, is simply a rule or formula or method that tells how to estimate the population parameter from the information provided by the sample at hand. A particular numerical value obtained by the estimator is known as **an estimate.** It should be noted that an estimator is random but an estimate is non random.

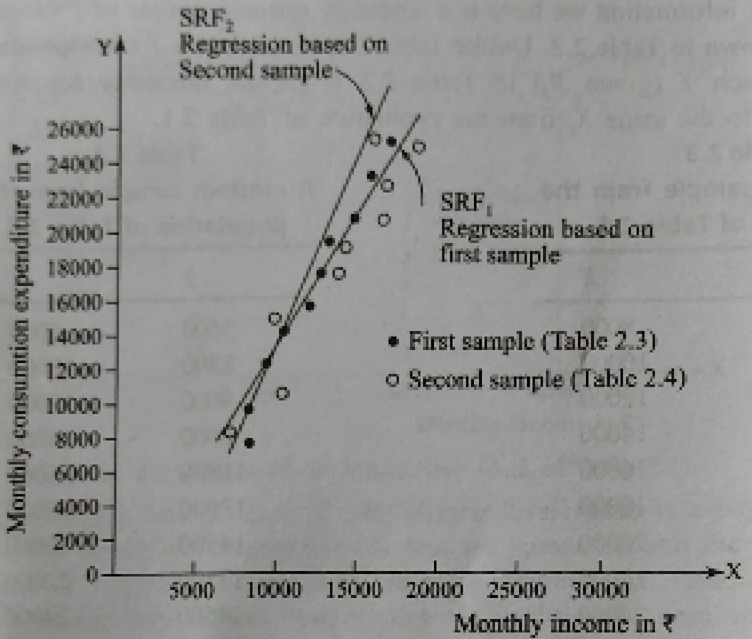


Fig. 2.2. Regression lines based on two different samples

Like the population regression function ($Y_i = E(Y/X_i) + u_i = \alpha + \beta X_i + u_i$) we can express the sample regression equation $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ in its stochastic form as follows:

$Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i$ where $\hat{u}_i$ denotes the (sample) residual term. Conceptually $\hat{u}_i$ is analogous to u_i and can be regarded as an estimate of u_i. It is introduced in the *SRF* for the same reasons as u_i was introduced in the *PRF*.

Now our primary objective in regression analysis is to estimate the *PRF*, $Y_i = \alpha + \beta X_i + u_i$ on the basis of the *SRF*, $Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i$. It should be noted that the estimate of the *PRF* based on the *SRF* is at best an approximate one (as sampling fluctuations exist). We have to develop procedures that tell us how to construct the *SRF* to mirror the *PRF* as faithfully as possible.

2.2. The Simple Linear Regression Model

Relationships suggested by economic theory are usually specified as exact or deterministic relationships between variables ; while on the other hand much stress is placed on the need for testing these economic theories. This implies a belief in the existence of stochastic function. The knowledge of econometrics tries to test these theoretical propositions in terms of stochastic variables. The simplest form of *stochastic relation* between two variables X and Y is called a simple linear regression model and is given by $Y_i = \alpha + \beta X_i + u_i$ for $i = 1, 2, ..., n$ where Y = dependent variable, X = explanatory variable (independent variable), u = stochastic disturbance term, α and β are two regression parameters whose values are to be determined on the basis of the given data on X and Y. The subscript 'i' refers to the i th observation, n = sample size or number of data points.

The stochastic nature of the regression model implies that for every value of X there is a whole probability distribution of values of Y. In other words, the value of Y can never be predicted exactly. This uncertainty concerning the value of Y arises because of the presence of the stochastic term 'u' which imparts randomness to Y.

2.2.1. Role of Random disturbance term in Econometric Model

We may ask why should we add an error term or random disturbance term u_i in an econometric model ?

The disturbance term u_i is a surrogate for all those variables that are omitted from the model but that collectively afect Y. We may give the following reasons for the insertion of the disturbance term in an econometric model.

(i) ***Omission of variables from the function :*** Suppose, in the model $Y = \alpha + \beta X + u$, the variable Y denotes the consumption expenditure and X denotes disposable income. But in reality X is not the only variable influencing Y. The family size, tastes of the family, spending habits and so on affect the variable Y. The error 'u' is a *catch-all* for the effects of all these variables, some of which may not even be quantifiable, and some of which may not even be identifiable. Therefore u_i may be used as a substitute for all the excluded or omitted variables from the model.

(ii) ***Unpredictable element of randomness in human responses :*** For instance, if Y = consumption expenditure of a household and X = disposable income of the household, there is an unpredictable element of randomness in each household's consumption. The household does not behave like a machine. In one month the people in the household are on a spending spree. In other month they are tightfisted.

(iii) ***Imperfect specification of the mathematical form of the model :*** We may have linearised a possibly nonlinear relationship between X and Y or we may have left out of the model some equations.

It is because the economic phenomena are much more complex than a single equation may reveal. For example, price determines and is determined by the quantity supplied (or quantity demanded) in the market. Under such circumstances if we attempt to study the phenomena with a single equation model, we are bound to commit an error. Thus the disturbance term represents such an error which may be due to imperfect specification of the term of the model, that is, of the number of equations.

(iv) ***Core variables versus peripheral variables :*** In consumption income relation (for instance) we may observe that besides income X_1, the number of children per family X_2, sex X_3, religion X_4, education X_5 and geographical region X_6 etc. also can affect consumption expenditure. But it is quite possible that the joint influence of all or some of these variables may be so small that as a practical matter it does not pay to introduce them into the model explicitly. However, their combined effect can be treated as a random variable u_i.

(v) ***Principle of Parsimony :*** Generally we would like to keep our regression model as simple as possible. If we can explain the behaviour of Y "substantially" with two or three explanatory variables and if our theory is not strong enough to suggest what other variables might be included, why introduce more variables ? In such cases u_i

represent all other variables. Of course, we should not exclude relevant and important variables just to keep the regression model simple.

(vi) ***Due to aggregation :*** We often use aggregate data (aggregate consumption, aggregate income), in which we add magnitudes referring to individuals whose behaviour is dissimilar. In this case we say that variables expressing individual peculiarities are missing. For example, in a production function for an industry we add together the factor inputs and outputs of dissimilar entrepreneurs. Changes in the distribution of total output among firms are important in the determination of total output. However, such distributional variables are often missing from the function. There are other types of aggregation which introduce error in the relationship.

(vii) ***Due to errors in measurement (Errors in variables) :*** Another justification for the insertion of a disturbance term is that it represents the measurement errors in the recording or processing of the data on X and Y. Thus the disturbance term reflects the errors in the observations.

For all these reasons, the stochastic disturbances u_i assume an extremely critical role in regression analysis.

2.3. Classical Linear Regression Model and its Assumptions

Let us consider an observed relation between two variables X and Y which is given by $Y_i = \alpha + \beta X_i + u_i$ for $i = 1, 2, ..., n$ where Y is the dependent variable, X is the independent variable, u is the disturbance term, the subscript 'i' denotes the item, n = no of observations, α and β are the two parameters whose values are to be estimated on the basis of the observed data on X and Y.

Now the model is called a Classical Linear Regression Model (CLRM) if the model satisfies the following properties (assumptions) :

Assumption 1 : u_i is a random variable which follows normal distribution.

Assumption 2 : $E(u_i) = 0$ for each i, $i = 1, 2, ..., n$. This means that the probability distribution of the disturbance term is such that its mean is zero.

Now $E(u_i) = 0$ implies $E(Y_i) = \alpha + \beta X_i$. [This can be shown as follows : Since $Y_i = \alpha + \beta X_i + u_i$. Now $E(Y_i) = E(\alpha + \beta X_i + u_i) = E(\alpha) + \beta E(X_i) + E(u_i) = \alpha + \beta X_i$, as $E(u_i) = 0$ and $E(X_i) = X_i$.]

But $\alpha + \beta X_i$ is the true value of Y_i. This means that expectation of observed value of the dependent variable is its true value. In other words the probability distribution of Y_i is centred around the true relationship.

Assumption 3 : Variance of each u_i is a constant and is independent of i, $i = 1, 2, ..., n$ and is denoted by σ_u^2 or simply σ^2.

i.e., $\text{Var}(u_i) = \sigma_u^2$ or σ^2

or, $E[u_i - E(u_i)]^2 = E(u_i)^2 = \sigma_u^2$, where $E(u_i) = 0$.

Assumptions 2 and 3 imply that the random variables $u_1, u_2, ..., u_n$ are identically distributed with the same mean (zero) and same variance (σ_u^2).

i.e., $u_i \sim ID(0, \sigma_u^2)$ for each i, $i = 1, 2, 3, ... n$.

Assumption 4 : The different error terms are independently distributed i.e., $E(u_i, u_j) = E(u_i).E(u_j)$.

Now $Cov\ (u_i, u_j) = 0$ for $i \neq j$

and $Cov\ (u_i, u_j) = \sigma_u^2$ for $i = j$ where $i, j = 1, 2, ..., n$.

Assumption 5 : The independent variable X is non-stochastic or non-random (which implies that X is not a random variable) and is measured without error, u_i is independent with explanatory variables.

i.e., $E\ (X_i, u_j) = X_i\ E(u_j) = 0$ for all $i, j = 1, 2, ..., n$.

The regression equation $Y = \alpha + \beta X + u$ along with the given five assumptions represents the Classical Linear Regression Model. The five assumptions have important roles to play in the sampling distributions of parameters α and β.

The effect of first three assumptions on the probability distribution of dependent variable Y can now be rationalised.

(i) In the equation $Y_i = \alpha + \beta X_i + u_i$, Y_i is a linear function of u_i. Since u_i is normally distributed, it follows that Y_i is also normally distributed.

(ii) $Y_i = \alpha + \beta X_i + u_i$

$$\begin{aligned} \therefore E(Y_i) &= E\ [\alpha + \beta X_i + u_i] && [\because E(\alpha) = \alpha,\ E\ (u_i) = 0] \\ &= \alpha + \beta X_i \end{aligned}$$

This means that the mean of Y_i is $\alpha + \beta X_i$

(iii) $$\text{Var}\ (Y_i) = E[Y_i - \overline{Y}]^2 = E[Y_i - E(Y_i)]^2 = E[\alpha + \beta X_i + u_i - (\alpha + \beta X_i)]^2$$

$$= E(u_i)^2 = \sigma_u^2 \left[\because E(u_i)^2 = \sigma_u^2\right]$$

Therefore, we say that variance of Y_i is σ_u^2.

Thus with the first three assumptions of u_i, we can indirectly say that Y_i is normally distributed with mean $(\alpha + \beta X_i)$ and variance σ_u^2.

Symbolically, $Y_i \sim N(\alpha + \beta X_i, \sigma_u^2)$ when $u_i \sim N(0, \sigma_u^2)$. This is illustrated in Fig. 2.3. Let $Y = \alpha + \beta X$ represent the population regression line. This regression line is unknown as we do not know the exact values of α and β. We have to estimate the values of α and β on the basis of sample data.

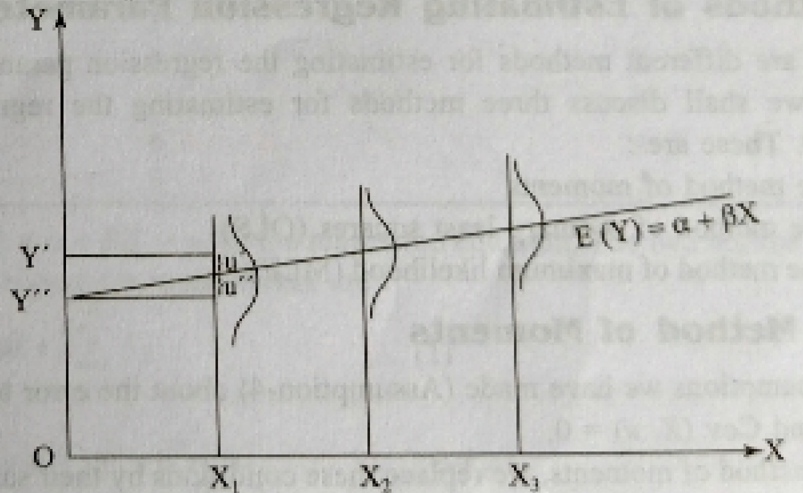


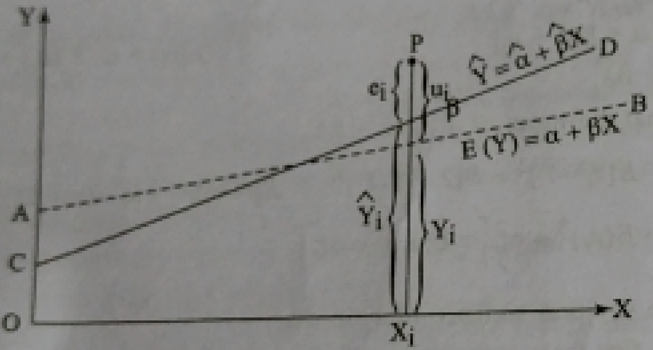
Fig. 2.3.

On substituting these estimated values in the *population regression line* we obtain *sample regression line*, which in turn serves as an estimate of the population regression line.

If the estimated values of α and β are given by $\hat{\alpha}$ and $\hat{\beta}$ respectively then sample regression line is given by $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$; $\hat{Y}_i$ represents the fitted values of actually observed values Y_i. We cannot expect the observed values (Y_i) to fit exactly on the sample regression line $(\hat{\alpha} + \hat{\beta}X_i)$.

Values of Y_i and $\hat{Y}_i$ will differ and this difference is a residual and is denoted by e_i where $e_i = Y_i - \hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i + e_i - \hat{\alpha} - \hat{\beta}X_i = e_i$. Thus the true relationship is given by $Y_i = \hat{\alpha} + \hat{\beta}X_i + e_i$. $Y_i = \hat{\alpha} + \hat{\beta}X_i + u_i$ and, the estimated relationship given by

The difference between these two relations along with the difference between residual and disturbance term has been illustrated in Fig. 2.4.



Y
P
$\hat{Y} = \hat{\alpha} + \hat{\beta}X$
D
e_i
u_i
β
B
$E(Y) = \alpha + \beta X$
A
$\hat{Y}_i$
Y_i
C
O
X
X_i

Fig. 2.4.

In Fig. 2.4. *AB* is the true regression line, *CD* is the estimated regression line. *P* represents one of the observations in the sample data. e_i differs from u_i because the true values of the parameters are different from their estimated values. In fact we can think the residual e_i as the estimate of the disturbance u_i.

2.4. Methods of Estimating Regression Parameters

There are different methods for estimating the regression parameters.

Now we shall discuss three methods for estimating the regression parameters (α and β). These are :

(i) The method of moments

(ii) The method of ordinary least squares (OLS)

(iii) The method of maximum likelihood (MLE).

2.5. The Method of Moments

The assumptions we have made (Assumption-4) about the error term 'u' imply that $E(u) = 0$ and Cov $(X, u) = 0$.

In the method of moments, we replace these conditions by their sample counterparts. Let $\hat{\alpha}$ and $\hat{\beta}$ be the estimators of α and β respectively.

Since $Y_i = \alpha + \beta X_i + u_i$ for $i = 1, 2, ..., n$ is the regression equation.

The sample counterpart of u_i is the estimated error (which is also called the residual) defined as $\hat{u}_i = Y_i - \hat{\alpha} - \hat{\beta} X_i$.

The two equations to determine $\hat{\alpha}$ and $\hat{\beta}$ are obtained by replacing population assumptions by their sample counterparts.

Population Assumption	Sample Counterpart
$E(u) = 0$	$\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i = 0$ or, $\sum_{i=1}^{n} \hat{u}_i = 0$
Cov $(X, u) = 0$	$\frac{1}{n}\sum_{i=1}^{n} X_i \hat{u}_i = 0$ or, $\sum_{i=1}^{n} X_i \hat{u}_i = 0$

Thus we get the two equations

$$\sum_{i=1}^{n} \hat{u}_i = 0 \text{ or, } \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i) = 0$$

and
$$\sum_{i=1}^{n} X_i \hat{u}_i = 0 \text{ or, } \sum_{i=1}^{n} X_i (Y_i - \hat{\alpha} - \hat{\beta} X_i) = 0$$

These equations can be written as :

$$\sum_{i=1}^{n} Y_i = n\hat{\alpha} + \hat{\beta} \sum X_i \qquad \text{.......} (1)$$

$$\sum_{i=1}^{n} X_i Y_i = \hat{\alpha} \sum X_i + \hat{\beta} \sum X_i^2 \qquad \text{.......} (2)$$

These two equations are called 'normal equations'. Solving these two equations we can get $\hat{\alpha}$ and $\hat{\beta}$.

Example : 2.1. Consider the data on advertising expenditures (X) and sales revenue (Y) for an athletic sports wear store for 5 months.

The observations are as follows :

Month	Sales Revenue, (Y) (in '000 ₹)	Advertising Expenditure, (X) (in '00 ₹)
1	3	1
2	4	2
3	2	3
4	6	4
5	8	5

Solution : Let $Y_i = \alpha + \beta X_i + u_i$ be the regression equation. The two normal equations for estimating the regression coefficients are :

$$\sum_{i=1}^{n} Y_i = n\hat{\alpha} + \hat{\beta} \sum_{i=1}^{n} X_i \qquad \text{......} (1)$$

$$\sum_{i=1}^{n} X_i Y_i = \hat{\alpha} \sum_{i=1}^{n} X_i + \hat{\beta} \sum_{i=1}^{n} X_i^2 \qquad \text{......} (2)$$

Calculations for $\hat{\alpha}$ and $\hat{\beta}$ (estimators of α & β)

Month	X_i	Y_i	X_i^2	X_iY_i	$\hat{u}_i = Y_i - \hat{\alpha} - \hat{\beta}X_i$
1	1	3	1	3	0.8
2	2	4	4	8	0.6
3	3	2	9	6	−2.6
4	4	6	16	24	0.2
5	5	8	25	40	1.0
Total	$\sum_{i=1}^{n} X_i = 15$	$\sum_{i=1}^{n} Y_i = 23$	$\sum_{i=1}^{n} X_i^2 = 55$	$\sum_{i=1}^{n} X_iY_i = 81$	$\Sigma \hat{u}_i = 0$

Here $n = 5$; Now from the two normal equations

we get, $5\hat{\alpha} + 15\hat{\beta} = 23$ ...... (1)

$15\hat{\alpha} + 55\hat{\beta} = 81$ .... (2)

Solving (1) and (2) by Cramer's rule we get,

$$\hat{\alpha} = \frac{\begin{vmatrix} 23 & 15 \\ 81 & 55 \end{vmatrix}}{\begin{vmatrix} 5 & 15 \\ 15 & 55 \end{vmatrix}} = \frac{1265-1215}{275-225} = \frac{50}{50} = 1 \text{ and } \hat{\beta} = \frac{\begin{vmatrix} 5 & 23 \\ 15 & 81 \end{vmatrix}}{\begin{vmatrix} 5 & 15 \\ 15 & 55 \end{vmatrix}} = \frac{405-345}{275-225} = \frac{60}{50} = 1{\cdot}2$$

Thus the estimated regression equation is

$$\hat{Y} = \hat{\alpha} + \hat{\beta}X \Rightarrow \hat{Y} = 1.0 + 1.2X.$$

The intercept 1.0 gives the value of Y when $X = 0$. This says that if advertising expenditures are zero, sales revenue will be ₹1000. The slope coefficient is 1.2. This says that if advertisement expenditure (X) is changed by 1 unit (₹100), sales revenue increases by 1.2 units (₹1,200 on an average). We have also shown the estimated errors or the residuals, given by

$\hat{u}_i = Y_i - 1.0 - 1.2X_i$, shown in the last column of the above table.

2.6. The Method of Ordinary Least Squares (OLS)

Let $Y_i = \alpha + \beta X_i + u_i$ be a two-variable linear regression model where Y is the dependent variable and X is the independent variable and u is the disturbance term. If the disturbance term u satisfies the following properties, then this model will be called a classical linear regression model (CLRM) :

(i) $E(u_i) = 0$ for each 'i', $i = 1, 2, 3, ..., n$

(ii) $E(u_i^2) = \sigma_u^2$ for each 'i'

(iii) $E(u_i, u_j) = 0$ for all $i, j, i \neq j$

(iv) $E(u_i, u_j) = \sigma_u^2$ for all $i = j$

(v) Independent variable X is non-stochastic.

The two parameters α and β of the regression equation can be obtained by the method of ordinary least squares (OLS). Let $\hat{\alpha}$ and $\hat{\beta}$ be the estimated values of α

and β. The estimated relation becomes $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$ and $e_i = Y_i - \hat{Y}_i$, is the residual term which shows the difference between the observed and estimated value.

The method of least squares consists in finding out those values of $\hat{\alpha}$ and $\hat{\beta}$ for which $\sum_{i=1}^{n} e_i^2$ is minimum. This means that we have to minimise $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta}X_i)^2$ through the choice of $\hat{\alpha}$ and $\hat{\beta}$. The necessary conditions of minimization requrie

$$\frac{\delta \sum_{i=1}^{n} e_i^2}{\delta \hat{\alpha}} = -2\sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0 \qquad ......(1)$$

and
$$\frac{\delta \sum_{i=1}^{n} e_i^2}{\delta \hat{\beta}} = -2\sum_{i=1}^{n} X_i(Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0 \qquad ......(2)$$

Simplifying equations (1) and (2) we get two normal equations

$$\sum_{i=1}^{n} Y_i = n\hat{\alpha} + \hat{\beta}\sum_{i=1}^{n} X_i \qquad ......(3)$$

$$\sum_{i=1}^{n} X_iY_i = \hat{\alpha}\sum_{i=1}^{n} X_i + \hat{\beta}\sum_{i=1}^{n} X_i^2 \qquad ......(4)$$

Now solving equations (3) and (4) by Cramer's rule we have,

$$\hat{\beta} = \frac{\begin{vmatrix} n & \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_iY_i \end{vmatrix}}{\begin{vmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 \end{vmatrix}} = \frac{n\sum_{i=1}^{n} X_iY_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} = \frac{Cov(X,Y)}{Var(X)}$$

or, $\hat{\beta} = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$, assuming $X_i - \bar{X} = x_i$ and $Y_i - \bar{Y} = y_i$

Again from equation (3) we get,

$$\sum_{i=1}^{n} Y_i = n\hat{\alpha} + \hat{\beta}\sum_{i=1}^{n} X_i$$

or, $$\sum_{i=1}^{n}\frac{Y_i}{n} = \hat{\alpha} + \hat{\beta}\sum_{i=1}^{n}\frac{X_i}{n} \quad \text{or, } \bar{Y} = \hat{\alpha} + \hat{\beta}\bar{X} \quad \therefore \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

2.6.1. Reverse Regression

By applying OLS method we have estimated the linear regression equation $Y_i = \alpha + \beta X_i + u_i$ where u_i satisfies all the properties of CLRM. The estimated regression equation becomes, $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$ where $\hat{\alpha}$ and $\hat{\beta}$ are the OLS estimators of α and β.

$$\text{Here } \hat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{Cov(X,Y)}{Var(X)} = \frac{r_{XY}\sigma_X \times \sigma_Y}{\sigma_X^2} = r_{XY}\frac{\sigma_Y}{\sigma_X}.$$

If we put $x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$, then we have

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \quad \text{and} \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}.$$

Here $\hat{\beta}$ is the estimated regression coefficient of Y on X. In this case the regression equation so defined is called the direct regression equation (of Y on X) where Y is the dependent variable and X is the independent variable. Sometimes we have to consider the regression equation of X on Y as well. This is called **reverse regression.**

The reverse regression is used in many cases. For instance, reverse regression has been advocated in the analysis of sex (or race) discrimination in salaries.

Suppose Y = salary and X = qualification and we are interested in determining if there is sex discrimination in salaries. We can ask :

1. Whether men and women with the same qualifications (value of X) are getting the same salaries (value of Y). This question is answered by the direct regression, i.e., regression of Y on X. Alternatively, we can ask :
2. Whether men and women with same salaries (value of Y) have the same qualifications (value of X).

This question is answered by the reverse regression, i.e., regression equation of X on Y.

For the reverse regression, the regression equation can be written as : $X_i = \alpha' + \beta' Y_i + v_i$ where v_i are the errors satisfying all the properties of CLRM. Here X is the dependent variable and Y is the independent variable.

The estimated relation becomes $\hat{X}_i = \hat{\alpha}' + \hat{\beta}' Y_i$ and $e_i' = X_i - \hat{X}_i$, which is the residual term, showing the difference between the observed and estimated relation. The method of least squares consists in finding out those values of $\hat{\alpha}'$ and $\hat{\beta}'$ for which $\sum_{i=1}^{n} e_i'^2$ is minimum. This means that we have to minimise

$\sum_{i=1}^{n} e_i'^2 = \sum_{i=1}^{n} (X_i - \hat{X}_i)^2 = \sum_{i=1}^{n} (X_i - \hat{\alpha}' - \hat{\beta}' Y_i)^2$ through the choice of $\hat{\alpha}'$ and $\hat{\beta}'$. The necessary conditions of minimisation require

$$\frac{\delta \sum_{i=1}^{n} e_i'^2}{\delta \hat{\alpha}'} = -2\sum_{i=1}^{n} (X_i - \hat{\alpha}' - \hat{\beta}' Y_i) = 0 \qquad \text{....... (1)}$$

and $$\frac{\delta \sum_{i=1}^{n} e_i'^2}{\delta \hat{\beta}'} = -2\sum_{i=1}^{n} Y_i (X_i - \hat{\alpha}' - \hat{\beta}' Y_i) = 0 \qquad \text{..... (2)}$$

Simplifying equations (1) and (2) we get two normal equations :

$$\sum_{i=1}^{n} X_i = n\hat{\alpha}' + \hat{\beta}' \sum_{i=1}^{n} Y_i \qquad \text{...... (3)}$$

$$\sum_{i=1}^{n} X_i Y_i = \hat{\alpha}' \sum_{i=1}^{n} Y_i + \hat{\beta}' \sum_{i=1}^{n} Y_i^2 \qquad \text{...... (4)}$$

Now solving equations (3) and (4) by Cramer's rule we have,

$$\hat{\beta}' = \frac{\begin{vmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} Y_i & \sum_{i=1}^{n} X_i Y_i \end{vmatrix}}{\begin{vmatrix} n & \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} Y_i & \sum_{i=1}^{n} Y_i^2 \end{vmatrix}} = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2}$$

$$= \frac{Cov(X,Y)}{Var(Y)} = \frac{r_{XY}\sigma_X\sigma_Y}{\sigma_Y^2} = r_{XY} \cdot \frac{\sigma_X}{\sigma_Y}$$

= Regression coefficient of X on Y.

If we put $x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$, then we can also express $\hat{\beta}'$ as follows :

$$\hat{\beta}' = \frac{Cov(X,Y)}{Var(Y)} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i^2}$$

Again from equation (3) we get

$$\sum_{i=1}^{n} X_i = n\hat{\alpha}' + \hat{\beta}'\sum_{i=1}^{n} Y_i$$

or, $$\sum_{i=1}^{n}\frac{X_i}{n} = \frac{n\hat{\alpha}'}{n} + \hat{\beta}'\sum_{i=1}^{n}\frac{Y_i}{n} \text{ or, } \bar{X} = \hat{\alpha}' + \hat{\beta}'\bar{Y} \text{ or, } \hat{\alpha}' = \bar{X} - \hat{\beta}'\bar{Y}.$$

It should be noted that $\hat{\beta}$ is the regression coefficient of Y on X and $\hat{\beta}'$ is the regression coefficient of X on Y.

Since $\hat{\beta} = r_{XY}\dfrac{\sigma_Y}{\sigma_X}$ and $\hat{\beta}' = r_{XY}\dfrac{\sigma_X}{\sigma_Y}$

$$\therefore\ \hat{\beta}\cdot\hat{\beta}' = r_{XY}\cdot\frac{\sigma_Y}{\sigma_X}\cdot r_{XY}\frac{\sigma_X}{\sigma_Y} = r_{XY}^2.$$

The two regression lines (Y on X and X on Y) will be different if $r_{XY}^2 < 1$ $\Rightarrow -1 < r_{XY} < 1$. The two regression lines will coincide if $r_{XY} = \pm 1$ and they will be perpendicular to each other if $r_{XY} = 0$.

Example 2.1.1. : We now consider a numerical example where we will fit both the regression (direct regression, i.e., Y on X and reverse regression i.e., X on Y) lines on the basis of the following data :

X:	10	7	10	5	8	8	6	7	9	10
Y:	11	10	12	6	10	7	9	10	11	10

where X = labour-hours of work, Y = output.

Solution : We know that the fitted direct regression equation (Y on X) is given by

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i \text{ where } \hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

and $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$ where $x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$. Conversely, the equation of the fitted reverse regression equation (X on Y) is given by,

$$\hat{X}_i = \hat{\alpha}' + \hat{\beta}'Y_i \text{ where } \hat{\beta}' = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i^2}, \text{ where } \hat{\alpha}' = \bar{X} - \hat{\beta}'\bar{Y} \text{ and } x_i = X_i - \bar{X},\ y_i = Y_i - \bar{Y}.$$

Calculations for direct and reverse regression lines

Observation	X_i	Y_i	$x_i = X_i - \bar{X}$	$y_i = Y_i - \bar{Y}$	x_iy_i	x_i^2	y_i^2
1	10	11	2	1.4	2.8	4	1.96
2	7	10	–1	0.4	–0.4	1	0.16
3	10	12	2	2.4	4.8	4	5.76
4	5	6	–3	–3.6	10.8	9	12.96
5	8	10	0	0.4	0	0	0.16
6	8	7	0	–2.6	0	0	6.76
$n=10$ 7	6	9	–2	–0.6	1.2	4	0.36
8	7	10	–1	0.4	–0.4	1	0.16
9	9	11	1	1.4	1.4	1	1.96
10	10	10	2	0.4	0.8	4	0.16
Total	$\Sigma X_i = 80$	$\Sigma Y_i = 96$	$\Sigma x_i = 0$	$\Sigma y_i = 0$	$\Sigma x_iy_i = 21$	$\Sigma x_i^2 = 28$	$\Sigma y_i^2 = 30.4$

$$\therefore \bar{X} = \frac{\Sigma X_i}{n} = \frac{80}{10} = 8, \quad \bar{Y} = \frac{\Sigma Y_i}{n} = \frac{96}{10} = 9.6$$

Now $\hat{\beta} = \dfrac{\Sigma x_i y_i}{\Sigma x_i^2} = \dfrac{21}{28} = 0.75$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 9.6 - 0.75 \times 8 = 3.6$$

$\therefore$ Estimated regression equation of Y on X (direct regression) is given by, $\hat{Y} = \hat{\alpha} + \hat{\beta}X$ or, $\hat{Y} = 3.6 + 0.75\,X$

Again, $\hat{\beta}' = \dfrac{\Sigma x_i y_i}{\Sigma y_i^2} = \dfrac{21}{30.4} = 0.690$

and $\hat{\alpha}' = \bar{X} - \hat{\alpha}'\bar{Y} = 8 - 0.690 \times 9.6 = 1.376$

$\therefore$ Estimated reverse regression equation (X on Y) is given by $\hat{X} = \hat{\alpha}' + \hat{\beta}'Y$

or, $\hat{X} = 1.376 + 0.690Y$

If should be noted that $\hat{\beta} = 0.75$ is the estimated regression coefficient of Y on X and $\hat{\beta}' = 0.690$ is the estimated regression coefficient of X on Y.

Since $\hat{\beta} \times \hat{\beta}' = r_{XY}^2$

$\therefore\ r_{XY}^2 = 0.75 \times 0.690 = 0.5175 \simeq 0.52$ and $r_{XY} = +\sqrt{0.5175} = 0.7193 \simeq 0.72.$

2.6.2. Scaling and Units of Measurement

In the regression analysis the units in which the regressand or the dependent variable (Y) and the regressor(s) are measured make difference in the regression results.

Suppose we like to regress Indian gross domestic savings (GDS) and gross domestic product (GDP), in rupees crore as well in rupees lakh measured in 1999-2000 prices. We also assume that in the regression of GDS on GDP one researcher uses data in rupees crore and another researcher uses data in rupees lakh. Now the natural questions : Will the regression results be the same in both cases ? Do the units in which the regressand and regressor(s) are measured make any difference in the regression results ? If so, what is the sensible course to follow in choosing units of measurement for regression analysis ? To answer these questions, let us proceed as follows :

Let $Y_i = \alpha + \beta X_i + u_i$ ...... (1)

where Y = GDS and X = GDP. Let us define

$$Y_i^* = W_1 Y_i \qquad \text{...... (2)}$$

and $X_i^* = W_2 X_i$ ...... (3)

where W_1 and W_2 are constants, called the scale factors ; W_1 may be equal to W_2 or may be different. If X_i and Y_i are measured in say rupees crore and we want to express them in rupees lakh, we will have $Y_i^* = 100\ Y_i$ and $X_i^* = 100\ X_i$, here $W_1 = W_2 = 100$.

Now consider the regression using Y_i^* and X_i^* variables :

$$Y_i^* = \alpha^* + \beta^*\ X_i^* + u_i^* \qquad \text{...... (4)}$$

where $Y_i^* = W_1 Y_i$, $X_i^* = W_2 X_i$ and $u_i^* = W_1 u_i$, or $W_I = W_2$

Now comparing equations (1) and (4) we can find out the relatioships between the following pairs :

1. $\hat{\alpha}$ and $\hat{\alpha}^*$
2. $\hat{\beta}$ and $\hat{\beta}^*$
3. Var $(\hat{\alpha})$ and Var $(\hat{\alpha}^*)$
4. Var $(\hat{\beta})$ and Var $(\hat{\beta}^*)$
5. $\hat{\sigma}_u^2$ and $\hat{\sigma}_u^2{}^*$
6. r_{XY}^2 and $r_{X^*Y^*}^2$

From the least squares theory we know that [applying OLS method on equation(1)]

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} \qquad \text{...... (5)}$$

$$\hat{\beta} = \frac{\Sigma x_i y_i}{\Sigma x_i^2} \qquad \text{...... (6) where } x_i = X_i - \overline{X} \text{ and } y_i = Y_i - \overline{Y}$$

$$\text{Var }(\hat{\alpha}) = \frac{\Sigma X_i^2}{n\Sigma x_i^2} \cdot \sigma_u^2 \qquad \text{...... (7)}$$

$$\text{Var }(\hat{\beta}) = \frac{\sigma_u^2}{\Sigma x_i^2} \qquad \text{...... (8)}$$

$$\text{and } \sigma_{\hat{u}}^2 = \frac{\Sigma \hat{u}_i^2}{n-2} \text{ or } \frac{\Sigma e_i^2}{n-2} \qquad \text{...... (9)}$$

Similarly, applying OLS method to equation (4) we obtain :

$$\hat{\alpha}^* = \bar{Y}^* - \hat{\beta}^* \bar{X}^* \quad \text{...... (10)}$$

$$\hat{\beta}^* = \frac{\Sigma x_i^* y_i^*}{\Sigma x_i^{*2}} \quad \text{...... (11) where } x_i^* = X_i^* - \bar{X}_i^* \text{ and } y_i^* = Y_i^* - \bar{Y}_i^*$$

$$\text{Var }(\hat{\alpha}^*) = \frac{\Sigma X_i^{*2}}{n\Sigma x_i^{*2}} \cdot \sigma_u^{*2} \quad \text{...... (12)}$$

$$\text{Var }(\hat{\beta}^*) = \frac{\sigma_u^{*2}}{\Sigma x_i^{*2}} \quad \text{...... (13)}$$

$$\hat{\sigma}_u^{*2} = \frac{\Sigma \hat{u}_i^{*2}}{n-2} \text{ or, } \frac{\Sigma e_i^{*2}}{n-2} \quad \text{...... (14)}$$

Thus we see that from model (1) $\hat{\alpha}$ and $\hat{\beta}$ are the OLS estimators of α and β and from model (2) $\hat{\alpha}^*$ and $\hat{\beta}^*$ are the OLS estimators of α^* and β^*. From the above results it is easy to establish relationship between two sets of parameters.

Since $Y_i^* = W_1 Y_i$ (or $y_i^* = w_1 y_i$) ; $X_i^* = W_2 X_i$ (or $x_i^* = w_2 x_i$) ; $u_i^* = w_1 u_i$; $\bar{Y}^* = W_1 \bar{Y}$ and $\bar{X}^* = W_2 \bar{X}$, we can easily verify that

$$\hat{\beta}_i^* = \left(\frac{W_1}{W_2}\right)\hat{\beta} \quad \text{...... (15)}$$

$$\hat{\alpha}^* = W_1 \hat{\alpha} \quad \text{...... (16)}$$

$$\hat{\sigma}_u^{*2} = W_1^2 \hat{\sigma}_u^2 \quad \text{..... (17)}$$

$$\text{Var }(\hat{\alpha}^*) = W_1^2 \text{ Var }(\hat{\alpha}) \quad \text{..... (18)}$$

$$\text{Var }(\hat{\beta}^*) = \left(\frac{W_1}{W_2}\right)^2 \cdot \text{Var }(\hat{\beta}) \quad \text{...... (19)}$$

$$r_{XY}^2 = r_{X^*Y^*}^2 \quad \text{...... (20)}$$

From the above results it is clear that from the regression results based on one scale of measurement, we can derive the results based on another scale of measurement once the scaling factors are kown. From the results given in (15) to (20) we can also derive some special cases. For instance, if the scaling factors are identical (i.e. $W_1 = W_2$) the slope coefficient and its standard error remain unaffected in going from the (Y_i, X_i) to the (Y_i^*, X_i^*) scale. However, the intercept and its standard error are both multiplied by W_1 (when $W_1 = W_2$). But if the X scale is not changed (i.e., $W_2 = 1$) and the Y scale is changed by the factor W_1 the slope as well as the intercept coefficients and their respective standard errors are all multiplied by the same W_1 factor. Finally, if the Y scale remains unchanged (i.e., $W_1 = 1$), but the X scale is changed by the factor W_2, the slope coefficient and its standard error are multiplied by the factor $\left(\frac{1}{W_2}\right)$ but the intercept coefficient and the standard error remain unaffected.

It should, however, be noted that the transformation of variables from the (Y, X) to the (Y^*, X^*) scale does not affect the properties of OLS estimators.

To illustrate the above theoretical results we consider an example, showing the relationship between GDS and GDP of India during the period 1951-52 to 2004-05.

The estimated regression equation of GDS on GDP (both GDS and GDP in rupees crore) is given by :

$$\hat{\text{GDS}}_t = -167423.37 + 0.36\text{GDP}_t \qquad \text{......} (21)$$

$$\text{SE}: \quad (17721.01) \quad (0.02) \qquad r^2 = 0.8891$$

Similarly, the estimated regression equation of GDS on GDP (both GDS and GDP in rupees lakh) is given by :

$$\hat{\text{GDS}}_t = -16742336.51 + 0.36\ \text{GDP}_t \qquad \text{......} (22)$$

$$\text{SE}: \quad (1772100.74) \quad (0.02) \qquad r^2 = 0.8891$$

Here we see that the intercept and its standard error is 100 times the corresponding values in the regression (21) [we should note that $W_1 = 100$ is going from crore to lakhs of rupees, i.e., 1 crore = 100 lakhs], but the slope coefficient as well as its standard error is unchanged, in accordance with the theory.

Now suppose we measure GDS in rupees crore and GDP in rupees lakh, the estimated regression equation becomes :

$$\hat{\text{GDS}}_t = -167423.37 + 0.0036\ \text{GDP}_t \qquad \text{......} (23)$$

$$\text{SE}: \quad (17721.01) \quad (0.0002), \quad r^2 = 0.8891$$

As expected, the slope coefficient as well as the standard error is $\left(\frac{1}{100}\right)$ its value in equation (21), since only X or GDP scale is changed.

If we express GDS in rupees lakh and GDP in rupees crore, the estimated regression equation becomes :

$$\hat{\text{GDS}}_t = -16742336.51 + 36.33\ \text{GDP}_t \qquad \text{......} (24)$$

$$\text{SE}: \quad (1772100.74) \quad (1.78), \qquad r^2 = 0.8891$$

Here we see that both the intercept and the slope coefficients as well as their respective standard errors are 100 times their values in equation (21) in accordance with our theoretical results.

If should be noted that the r^2 value remains the same in all the cases as it is invariant to changes in the unit of measurement and scales.

2.7. Estimation of a Function whose Intercept is Zero

In some cases economic theory postulates relationships which have a zero intercept, that is, they pass through the origin of the XY plane. [For example, long run consumption function of the form $C = bY$, where b = APC = MPC, C = consumption expenditure, Y = income].

In this event we should estimate the function $Y = \alpha + \beta X + u$, imposing the restriction $\alpha = 0$. The formula for the estimation of $\hat{\beta}$ then becomes $\hat{\beta} = \dfrac{\sum_{i=1}^{n} X_iY_i}{\sum_{i=1}^{n} X_i^2}$ which involves the actual values of the variables, and not their deviations, as in the case of unrestricted value of α.

Proof : We want to fit the line $Y_i = \alpha + \beta X_i + u_i$, subject to the restriction $\alpha = 0$. To estimate β, the problem is put in a form of restricted minimization problem and then Lagrange method is applied.

Now we have to minimize

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}X_i)^2$$

subject to $\hat{\alpha} = 0$.

The (Lagrange) composite function then becomes $L = \sum_{i=1}^{n}\left(Y_i - \hat{\alpha} - \hat{\beta}X_i\right)^2 - \lambda\hat{\alpha}$ where λ is the Lagrange multiplier. Now we have to minimize 'L' with respect to $\hat{\alpha}$, $\hat{\beta}$ and λ. First order conditions of minimization require :

$$\frac{\partial L}{\partial \hat{\alpha}} = -2\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}X_i) - \lambda = 0 \qquad \text{......(1)}$$

$$\frac{\partial L}{\partial \hat{\beta}} = -2\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}X_i)X_i = 0 \qquad \text{...... (2)}$$

$$\frac{\partial L}{\partial \lambda} = -\hat{\alpha} = 0 \qquad \text{...... (3)}$$

Now substituting (3) in (2) and rearranging we get,

$$-2\sum_{i=1}^{n} X_i(Y_i - \hat{\beta}X_i) = 0$$

or, $\sum_{i=1}^{n} X_iY_i - \hat{\beta}\sum_{i=1}^{n} X_i^2 = 0 \quad \therefore \hat{\beta} = \dfrac{\sum_{i=1}^{n} X_iY_i}{\sum_{i=1}^{n} X_i^2}$

In this case (i) $\hat{\sigma}_u^2 = \Sigma e_i^2/(n-1)$; (ii) $SE(\hat{\beta}) = \sqrt{\hat{\sigma}_u^2/\Sigma X_i^2}$; (iii) $R^2 = 1 - \Sigma e_i^2/\Sigma Y_i^2$.

2.8. Estimation of Elasticities from an Estimated Regression Line

The estimated regression equation is $\hat{Y}=\hat{\alpha}+\hat{\beta}X$ whose intercept is $\hat{\alpha}$ and slope $\hat{\beta}$ is the derivative of $\hat{Y}$ with respect to X i.e., $\hat{\beta}=\frac{d\hat{Y}}{dX}$ which shows the rate of change in $\hat{Y}$ as X changes by a very small amount. It should be clear that if the estimated function is a linear demand or supply function, the coefficient $\hat{\beta}$ is not the price elasticity, but a component of the elasticity, which is defined by the formula

$$\eta_p=\frac{dY/Y}{dX/X}=\frac{X}{Y}\cdot\frac{dY}{dX}.$$

where η_p = price elasticity, Y = quantity (demanded or supplied), X = price. Clearly $\hat{\beta}$ is the component $\frac{dY}{dX}$. From an estimated function we can obtain an average elasticity $\eta_p=\hat{\beta}\cdot\frac{\bar{X}}{\bar{\hat{Y}}}$ where $\bar{X}$ = the average price in the sample. $\bar{\hat{Y}}$ = average regressed value of the quantity, i.e., the mean value as estimated from the regression $\hat{Y}_i$, $\bar{Y}$ = average value of the quantity in the sample. It should be noted that $\bar{\hat{Y}}=\bar{Y}$. Since $\hat{Y}=\hat{\alpha}+\hat{\beta}X$.

$$\therefore \bar{\hat{Y}}=\hat{\alpha}+\hat{\beta}\bar{X}=(\bar{Y}-\hat{\beta}\bar{X})+\hat{\beta}\bar{X}=\bar{Y}.$$

In particular if $Y_i=\alpha+\beta X_i$ is the regression equation, then the estimated average elasticity $\hat{\eta}_p=\hat{\beta}\frac{\bar{X}}{\bar{Y}}$ where $\bar{Y}=\hat{\alpha}+\hat{\beta}\bar{X}$.

Now substituting for $\bar{Y}$ in the expression of elasticites, we obtain, $\hat{\eta}_p=\frac{\hat{\beta}\bar{X}}{\hat{\alpha}+\hat{\beta}\bar{X}}$.

If the function $Y=\alpha+\beta X$ represents a supply function with $\hat{\beta}>0$, it follows that

(i) the supply function will be elastic ($\eta_p>1$) if $\hat{\alpha}$ is negative ($\hat{\alpha}<0$).

(ii) the supply function will be inelastic ($\eta_p<1$) if $\hat{\alpha}>0$.

(iii) the supply function will have unitary elasticity ($\eta_p=1$) if $\hat{\alpha}=0$.

Thus the elasticity of a supply curve (with positive slope) depends on the sign of the constant intercept $\hat{\alpha}$.

Example 2.2. The following table includes the price and quantity demanded of the product of a monopolist over a six year period.

Year :	2014	2015	2016	2017	2018	2019
Quantity('000 Kg.) :	8	3	4	7	8	0
Price ('00 ₹)	2	4	3	1	3	5

(a) Estimate the demand function, assuming a linear demand function. Comment on the values of the estimated coefficients ($\hat{\alpha}$ and $\hat{\beta}$) on the basis of economic theory.

(b) Estimate the average elasticity of demand.

(c) Estimate the elasticity of demand at the price 4.

(d) Forecast the level of demand if price rises to 5. Comment on your forecast.

Solution : (a) Let $Y_i = \alpha + \beta X_i$ for $i = 1, 2, ...6$, be the linear demand function. By the OLS method we can get the estimators of α and β. Here Y = demand, X = price, α, β are two paramaters. Theoretically we may assume $\alpha > 0$, $\beta < 0$. By OLS method $\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$, where $x_i = X_i - \bar{X}$, $y_i = Y_i - \bar{Y}$ and $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$, $\bar{X} = \sum X_i / n$, $\bar{Y} = \sum Y_i / n$.

Calculations for the parameters (α, β)

Year (n)	quantity (1000 Kg.) Y_i	price ('00 ₹) X_i	$y_i = Y_i - \bar{Y}$	$x_i = X_i - \bar{X}$	$x_i y_i$	x_i^2
2014 (1)	8	2	3	− 1	−3	1
2015 (2)	3	4	− 2	1	−2	1
2016 (3)	4	3	− 1	0	0	0
2017 (4)	7	1	+ 2	− 2	− 4	4
2018 (5)	8	3	3	0	0	0
2019 (6)	0	5	− 5	2	− 10	4
$n = 6$ Total	$\Sigma Y_i = 30$	$\Sigma X_i = 18$	$\Sigma y_i = 0$	$\Sigma x_i = 0$	$\Sigma x_i y_i = -19$	$\Sigma x_i^2 = 10$

$$\therefore \bar{Y} = \frac{\sum Y_i}{n} = \frac{30}{6} = 5, \quad \bar{X} = \frac{\sum X_i}{n} = \frac{18}{6} = 3$$

$$\text{Now } \hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{-19}{10} = -1.9$$

and $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 5 - (-1.9) \times 3 = 5 + 5.7 = 10.7$

Thus the OLS estimators of α and β are $\hat{\alpha} = 10.7 > 0$ and $\hat{\beta} = -1.9 < 0$.

Therefore the estimated demand function is $\hat{Y} = \hat{\alpha} + \hat{\beta}X$ or $\hat{Y} = Y = 10.7 - 1.9X$.

This is consistent with the theory where we assume $\alpha > 0$ and $\beta < 0$. This clearly shows that there exists an inverse relation between price and demand (i.e., the law of demand holds true).

(b) The average elasticity (price elasticity of demand) is given by,

$\hat{\eta}_p = \hat{\beta} \cdot \frac{\bar{X}}{\bar{Y}} = -1.9 \times \frac{3}{5} = -1.14$ or, $|\hat{\eta}_p| = 1.14 > 1$.

This means that the demand function shows an elastic demand.

(c) We have to estimate η_p (price elasticity of demand), from the estimated relation $\hat{Y} = Y = 10.7 - 1.9X$ when price $= X = 4$.

If $X = 4$, $Y = 10.7 - 1.9 \times 4 = 10.7 - 7.6 = 3.1$

Now η_p (at $X = 4$) $= \frac{X}{Y} \cdot \frac{d\hat{Y}}{dX} = \frac{4}{3.1} \times -1.9 = -2.45$

$\therefore |\eta_p|$ at $X = 4$ is $2.45 > 1$.

This implies that the demand is elastic demand.

(d) We have to forecast the level of demand when price rises to 5 (i.e., when $X = 5$). From the estimated demand function

we get, $\hat{Y} = 10.7 - 1.9X$.

When $X = 4$, $\hat{Y} = 10.7 - 1.9 \times 4 = 3.1$

If now $X = 5$, $\hat{Y} = 10.7 - 1.9 \times 5 = 10.7 - 9.5 = 1.2$

This means that if price increases from 4 to 5, demand decreases from 3.1 to 1.2. This clearly shows that as price rises, demand decreases.

Example 2.3. The following table shows ten pairs of observations on X (price) and Y (quantity supplied)

No. of observations (n) :	1	2	3	4	5	6	7	8	9	10	11	12
Quantity (Y) (in tons) :	69	76	52	56	57	77	58	55	67	53	72	64
Price (X) (in '00 ₹) :	9	12	6	10	9	10	7	8	12	6	11	8

(*a*) Assuming a linear supply function estimate the supply function. Comment on the values of the estimated coefficients ($\hat{\alpha}$ and $\hat{\beta}$) on the basis of economic theory.

(*b*) Estimate the average price elasticity of supply.

(*c*) Estimate the elasticity of supply at the price 6.

(*d*) Forecast the level of supply if price rises to 8.

Solution : Let $Y_i = \alpha + \beta X_i$ for $i = 1, 2, ..., 12$ be the linear supply function. By OLS method we can get the estimators of α and β. Here, Y = supply, X = price, α, β are two parameters. Theoretically we may assume $\alpha \gtreqless 0$ and $\beta > 0$. By OLS method we may get,

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} \text{ where } \quad x_i = \bar{X}_i - \bar{X} \quad y_i = Y_i - \bar{Y}$$

$$\text{and } \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \qquad \bar{X} = \sum X_i / n,\ \bar{Y} = \sum Y_i / n$$

Calculations for the OLS estimators of parameters (α, β)

Observations n	Y_i Quantity (in tons)	X_i Price (in '00 ₹.)	x_i $= X_i - \bar{X}$	y_i $= Y_i - \bar{Y}$	$x_i y_i$	x_i^2
1	69	9	0	6	0	0
2	76	12	3	13	39	9
3	52	6	− 3	− 11	33	9
4	56	10	1	− 7	− 7	1
5	57	9	0	− 6	0	0
$n = 12$ 6	77	10	1	14	14	1
7	58	7	− 2	− 5	10	4
8	55	8	− 1	− 8	8	1
9	67	12	3	4	12	9
10	53	6	− 3	− 10	30	9
11	72	11	2	9	18	4
12	64	8	− 1	1	− 1	1
Total	$\Sigma Y_i = 756$	$\Sigma X_i = 108$	$\Sigma x_i = 0$	$\Sigma y_i = 0$	$\Sigma x_i y_i = 156$	$\Sigma x_i^2 = 48$

$$\therefore \overline{X} = \frac{\sum X_i}{n} = \frac{108}{12} = 9\ ,\ \overline{Y} = \frac{\sum Y_i}{n} = \frac{756}{12} = 63$$

(a) Now the OLS estimators of the regression parameters α and β are given by,

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{156}{48} = 3.25$$

and $\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} = 63 - 3.25 \times 9 = 63 - 29.25 = 33.75.$

Thus the estimated supply function is $\hat{Y} = \hat{\alpha} + \hat{\beta}X$ or, $\hat{Y} = 33.75 + 3.25X.$

Here we see that $\hat{\alpha} = 33.75 > 9$ and $\hat{\beta} = 3.25 > 0$. This means that there is a direct (positive) relation between supply and price. The intercept of the supply function is positive here. Hence our results are consistent with the theory.

(b) Average price elasticity of supply is given by, $\eta_p = \hat{\beta} \cdot \frac{\overline{X}}{\overline{Y}} = 3.25 \times \frac{9}{63} = 0.46 < 1$

This shows that at the average price the supply is price-inelastic.

(c) We have to find price elasticity of supply at price 6.

Since the estimated supply function is

$\hat{Y} = \hat{\alpha} + \hat{\beta}X$ or, $\hat{Y} = 33.75 + 3.25X$

Now if $X = 6$, $Y = 33.75 + 3.25 \times 6 = 33.75 + 19.5 = 53.25$

Now by definition price elasticity of supply, $\eta_p = \frac{X}{Y} \cdot \frac{d\hat{Y}}{dX} = \frac{6}{53.25} \times 3.25 = 0.366$

Thus $\eta_p = 0.366$ when $X = 6$.

(d) From the estimated supply function we see that $Y = 33.75 + 3.25X$.

When $X = 6$, $Y = 53.25$

If now price increases to 8 i.e., if $X = 8$

then $Y = 33.75 + 3.25 \times 8 = 33.75 + 26 = 59.75$.

This means that when $X = 6$, $Y = 53.25$

and when $X = 8$, $Y = 59.75$

Thus we may forecast that as price increases, supply will also increase.

2.9. Properties of Least Squares Estimators

The least squares estimates are called BLUE (best, linear, unbiased, estimates) provided that the random term u satisfies some general assumptions, namely that the u has zero mean and constant variance. This proposition, together with the set of conditions under which it is true, is known as **Gauss Markov Least-Squares Theorem.**

The OLS estimators possess three properties ; They are linear, unbiased and have the smallest variance (compared to other linear unbiased estimators). Thus the OLS estimators are BLUE.

1. The property of linearity

The least-squares estimates $\hat{\alpha}$ and $\hat{\beta}$ are linear functions of the observed sample values Y_i.

Since $\hat{\beta} = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$

and $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$

Now $\hat{\beta} = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$

$$= \frac{\sum_{i=1}^{n} Y_i(X_i - \bar{X}) - \bar{Y}\sum_{i=1}^{n}(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} Y_i(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad \left[\because \sum_{i=1}^{n}(X_i - \bar{X}) = 0\right]$$

$\therefore \hat{\beta} = \dfrac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}$, where $x_i = (X_i - \bar{X})$

Let us suppose that $\dfrac{x_i}{\sum_{i=1}^{n} x_i^2} = K_i (i = 1, 2, ..., n)$

$\therefore \hat{\beta} = \sum_{i=1}^{n} K_i Y_i$.

This shows that $\hat{\beta}$ is a linear function of Y_i.

Similarly, $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = \frac{1}{n}\sum_{i=1}^{n} Y_i - \bar{X}\sum_{i=1}^{n} K_i Y_i$ $\left[\text{where } \hat{\beta} = \sum_{i=1}^{n} K_i Y_i\right]$

$\therefore \hat{\alpha} = \sum_{i=1}^{n}\left[\frac{1}{n} - \bar{X}K_i\right]Y_i$.

This shows that $\hat{\alpha}$ is a linear function of Y_i.

Thus both $\hat{\alpha}$ and $\hat{\beta}$ are expressed as linear functions of the Y's.

2. The property of unbiasedness

The means of $\hat{\alpha}$ $[E(\hat{\alpha})]$ and $\hat{\beta}$ $[E(\hat{\beta})]$ can be obtained as follows :

$$\text{Since } \hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$= \frac{\sum_{i=1}^{n} Y_i(X_i - \bar{X}) - \bar{Y}\sum_{i=1}^{n}(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2} \text{ where } \sum_{i=1}^{n}(X_i - \bar{X}) = 0$$

and $x_i = X_i - \bar{X}$ for $i = 1, 2, ..., n$

$$\therefore \hat{\beta} = \sum_{i=1}^{n} K_i Y_i \text{ where } K_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2} \text{ and } \hat{\alpha} = \sum_{i=1}^{n}\left[\frac{1}{n} - \bar{X}K_i\right]Y_i$$

Now $\hat{\beta} = \sum_{i=1}^{n} K_i Y_i$. We now put $Y_i = \alpha + \beta X_i + u_i$

$$\therefore \hat{\beta} = \sum_{i=1}^{n} K_i(\alpha + \beta X_i + u_i) = \alpha\sum_{i=1}^{n} K_i + \beta\sum_{i=1}^{n} K_i X_i + \sum_{i=1}^{n} K_i u_i$$

$$\text{Since } K_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2} \quad \therefore \sum_{i=1}^{n} K_i = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} = 0, \text{ as } \sum_{i=1}^{n} x_i = 0$$

$$\text{and } \sum_{i=1}^{n} K_i X_i = \sum_{i=1}^{n} K_i(x_i + \bar{X}) \text{ where } x_i = X_i - \bar{X} \quad \therefore X_i = x_i + \bar{X}$$

$$= \sum_{i=1}^{n} K_i x_i + \bar{X}\sum_{i=1}^{n} K_i = \sum_{i=1}^{n} K_i x_i \quad \left[\because \sum_{i=1}^{n} K_i = 0\right]$$

$$\text{Now } \sum_{i=1}^{n} K_i x_i = \sum_{i=1}^{n} x_i x_i \Big/ \sum_{i=1}^{n} x_i^2 = \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^2} = 1, \text{ as } K_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2}$$

If we put $\sum_{i=1}^{n} K_i = 0$ and $\sum_{i=1}^{n} K_i X_i = 1$ in the expression of

$\hat{\beta} = \alpha\sum_{i=1}^{n} K_i + \beta\sum_{i=1}^{n} K_i X_i + \sum_{i=1}^{n} K_i u_i$, we get, $\hat{\beta} = \beta.1 + \sum_{i=1}^{n} K_i u_i$

Now mean of $\hat{\beta} = E(\hat{\beta}) = E(\beta) + \sum_{i=1}^{n} K_i E(u_i) = \beta$ $[\because E(u_i) = 0]$

Thus we have $E(\hat{\beta}) = \beta$ i.e., mean of $\hat{\beta}$ is β.

Similarly, $\hat{\alpha} = \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{X}K_i\right)Y_i$

$$= \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{X}K_i\right)(\alpha + \beta X_i + u_i) \quad [\because Y_i = \alpha + \beta X_i + u_i]$$

$$= \sum_{i=1}^{n}\frac{1}{n}\alpha + \beta\cdot\frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n}\sum_{i=1}^{n} u_i - \alpha\bar{X}\sum_{i=1}^{n} K_i - \beta\bar{X}\sum_{i=1}^{n} K_i X_i - \bar{X}\sum_{i=1}^{n} K_i u_i$$

Since $\sum_{i=1}^{n} K_i = 0$, $\sum_{i=1}^{n} K_i X_i = 1$, $\sum_{i=1}^{n} 1 = n$ we have,

$\hat{\alpha} = \alpha + \beta\bar{X} + \frac{1}{n}\sum_{i=1}^{n} u_i - \beta\bar{X} - \bar{X}\sum_{i=1}^{n} K_i u_i$

or, $\hat{\alpha} = \alpha + \frac{1}{n}\sum_{i=1}^{n} u_i - \bar{X}\sum_{i=1}^{n} K_i u_i$ or, $E(\hat{\alpha}) = E(\alpha) + \frac{1}{n}\sum_{i=1}^{n} E(u_i) - \bar{X}\sum_{i=1}^{n} K_i E(u_i)$

$\therefore E(\hat{\alpha}) = \alpha$, as $E(u_i) = 0$

This shows that mean of $\hat{\alpha}$ is α

Thus, it is proved that $\hat{\alpha}$ and $\hat{\beta}$ are unbiased estimators of α and β.

3. The minimum variance property

In this property we shall prove the Gauss Markov Theorem, which states that the least squares estimates are best (have the smallest variance) as compared with any other linear unbiased estimator obtained from other econometric methods.

First we have to find *Var* ($\hat{\beta}$) and *Var* ($\hat{\alpha}$) and then we have to prove the minimum variance property.

Variance of $\hat{\beta} = Var(\hat{\beta}) = E\left[\hat{\beta} - E(\hat{\beta})\right]^2 = E[\hat{\beta} - \beta]^2$ as $E(\hat{\beta}) = \beta$.

Since $\hat{\beta} = \beta + \sum_{i=1}^{n} K_i u_i$ (see property 2)

$$\therefore\ \hat{\beta} - \beta = \sum_{i=1}^{n} K_i u_i$$

$$\text{or, } (\hat{\beta} - \beta)^2 = \left[\sum_{i=1}^{n} K_i u_i\right]^2 \quad \text{or, } E[\hat{\beta} - \beta]^2 = E\left[\sum_{i=1}^{n} K_i u_i\right]^2$$

$$\text{or, } Var(\hat{\beta}) = E\left[\sum_{i=1}^{n} K_i^2 u_i^2 + 2\sum_{i \neq j}\sum K_i K_j u_i u_j\right]$$

$$= \sum_{i=1}^{n} K_i^2 E(u_i^2) + 2\sum_{i \neq j}\sum K_i K_j E(u_i u_j)$$

$$= \sum_{i=1}^{n} K_i^2 E(u_i^2), \quad [\because E(u_i u_j) = 0] \text{ for } i \neq j$$

$$= \sum_{i=1}^{n} K_i^2 \cdot \sigma_u^2 \quad [\because E(u_i^2) = \sigma_u^2]$$

$$= \frac{\sum_{i=1}^{n} x_i^2}{\left(\sum_{i=1}^{n} x_i^2\right)^2} \cdot \sigma_u^2 \qquad \left[\because K_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2}\right]$$

$$= \frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2} \qquad \therefore Var(\hat{\beta}) = \frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2}$$

Similarly, Variance of $\hat{\alpha} = Var(\hat{\alpha})$.

Since $\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}$ (see property 1)

Substituting $\hat{\beta} = \sum_{i=1}^{n} K_i Y_i$ we obtain $\hat{\alpha} = \overline{Y} - \overline{X}\sum_{i=1}^{n} K_i Y_i$

$$= \frac{\sum_{i=1}^{n} Y_i}{n} - \overline{X}\sum_{i=1}^{n} K_i Y_i = \sum_{i=1}^{n}\left(\frac{1}{n} - \overline{X}K_i\right)Y_i$$

Now, $Var(\hat{\alpha}) = Var\left[\sum\left(\frac{1}{n}-\bar{X}K_i\right)Y_i\right] = \sum_{i=1}^{n}\left(\frac{1}{n}-\bar{X}K_i\right)^2 \cdot Var\,(Y_i)$

Since $Var\,(Y_i) = \sigma_u^2$

$$\therefore\ Var\,(\hat{\alpha}) = \sum_{i=1}^{n}\left(\frac{1}{n}-\bar{X}K_i\right)^2 \cdot \sigma_u^2$$

$$= \sigma_u^2\sum_{i=1}^{n}\left(\frac{1}{n}-\bar{X}K_i\right)^2 = \sigma_u^2\sum_{i=1}^{n}\left(\frac{1}{n^2}-2\cdot\frac{1}{n}\cdot\bar{X}K_i+\bar{X}^2K_i^2\right)$$

$$= \sigma_u^2\left[\frac{1}{n}+\frac{\bar{X}^2}{\sum_{i=1}^{n}x_i^2}\right]\left[\because \sum_{i=1}^{n}K_i = 0 \text{ and } \sum_{i=1}^{n}K_i^2 = \frac{1}{\sum_{i=1}^{n}x_i^2} \text{ where} K_i = \frac{x_i}{\sum_{i=1}^{n}x_i^2}\right]$$

$$= \sigma_u^2\left[\frac{\sum_{i=1}^{n}x_i^2+n\bar{X}^2}{n\sum_{i=1}^{n}x_i^2}\right] = \sigma_u^2\cdot\frac{\sum_{i=1}^{n}X_i^2}{n\sum_{i=1}^{n}x_i^2}$$

$$\therefore Var\,(\hat{\alpha}) = \sigma_u^2\sum_{i=1}^{n}X_i^2 \Big/ n\sum_{i=1}^{n}x_i^2$$

$$\left[\because \sum_{i=1}^{n}x_i^2+n\bar{X}^2 = \sum_{i=1}^{n}(X_i-\bar{X})^2+n\bar{X}^2 = \sum_{i=1}^{n}X_i^2-2\bar{X}\sum_{i=1}^{n}X_i+n\bar{X}^2+n\bar{X}^2\right.$$

$$\left. = \sum_{i=1}^{n}X_i^2-2n\bar{X}^2+2n\bar{X}^2 = \sum_{i=1}^{n}X_i^2\right]$$

Case (a) $\hat{\beta}$ has the least variance :

We know that $Var(\hat{\beta}) = \sigma_u^2\Big/\sum_{i=1}^{n}x_i^2$.

Now we want to prove that any other linear unbiased estimate of the true parameter, for example β^*, obtained from any other econometric method, has a bigger variance than the least squares estimate $\hat{\beta}$. Thus we have to prove that $Var\,(\hat{\beta}) < Var\,(\beta^*)$

Proof : The new estimator β^* is by assumption a linear combination of the Y_i's, a weighted sum of the sample values Y_i, the weights $K_i\left(= x_i\Big/\sum_{i=1}^{n}x_i^2\right)$ being different from the weights of the least-squares estimates.

For example, let us assume $\beta^* = \sum_{i=1}^{n} C_i Y_i$ where $C_i = K_i + d_i$, d_i is an arbitrary set of weights similar (but not the same) to the K_i's.

Let us put $Y_i = \alpha + \beta X_i + u_i$ in the expression of β^* and we obtain

$$\beta^* = \sum_{i=1}^{n} C_i(\alpha + \beta X_i + u_i) = \sum_{i=1}^{n} (\alpha C_i + \beta C_i X_i + C_i u_i)$$

It is assumed that like $\hat{\beta}$, β^* is also an unbiased estimator of β, i.e., $E(\beta^*) = \beta$.

Now $E(\beta^*) = E\left[\sum_{i=1}^{n} (\alpha C_i + \beta C_i X_i + C_i u_i)\right]$

$$\therefore E(\beta^*) = E\left[\alpha \sum_{i=1}^{n} C_i + \beta \sum_{i=1}^{n} C_i X_i + \sum_{i=1}^{n} C_i u_i\right]$$

Now $E(\beta^*) = \beta$ if, and only if

$$\sum_{i=1}^{n} C_i = 0, \quad \sum_{i=1}^{n} C_i X_i = 1 \text{ and } \sum_{i=1}^{n} C_i u_i = 0$$

But $\sum_{i=1}^{n} C_i = 0$ implies $\sum_{i=1}^{n} d_i = 0$, because

$$\sum_{i=1}^{n} C_i = \sum_{i=1}^{n} (K_i + d_i) = \sum_{i=1}^{n} K_i + \sum_{i=1}^{n} d_i \text{ and } \sum_{i=1}^{n} K_i = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} = 0 \quad \left(\text{as } \sum_{i=1}^{n} x_i = 0\right)$$

i.e., $\sum_{i=1}^{n} C_i = \sum_{i=1}^{n} d_i$. Therefore if $\sum_{i=1}^{n} C_i = 0$, then $\sum_{i=1}^{n} d_i = 0$.

Similarly, $\sum_{i=1}^{n} C_i X_i = 1$ requires $\sum_{i=1}^{n} d_i X_i = 0$,

since $\sum_{i=1}^{n} C_i X_i = \sum_{i=1}^{n} (K_i + d_i) X_i = \sum_{i=1}^{n} K_i X_i + \sum_{i=1}^{n} d_i X_i$

Given that $\sum_{i=1}^{n} K_i X_i = 1, \sum_{i=1}^{n} C_i X_i = 1$ if $\sum_{i=1}^{n} d_i X_i = 0$

Thus β^* will be a linear unbiased estimate of β (with weights $C_i = K_i + d_i$) if

$$\sum_{i=1}^{n} C_i = 0, \quad \sum_{i=1}^{n} d_i = 0, \quad \sum_{i=1}^{n} C_i X_i = 1 \text{ and } \sum_{i=1}^{n} d_i X_i = 0$$

Since (from property 1) we know that,

$$\hat{\beta} = \sum_{i=1}^{n} K_i Y_i \text{ and } Var(\hat{\beta}) = Var\left(\sum_{i=1}^{n} K_i Y_i\right)$$

$$= \sum_{i=1}^{n} Var(K_i Y_i) = \sum_{i=1}^{n} K_i^2 Var(Y_i) = \sum_{i=1}^{n} K_i^2 \sigma_u^2$$

$$\left[\because Var(Y_i) = E[Y_i - E(Y_i)]^2 = \sigma_u^2 \text{ where } Y_i = \alpha + \beta X_i + u_i\right.$$

$$\therefore E(Y_i) = E(\alpha + \beta X_i) + E(u_i)$$

$$= \alpha + \beta X_i, \text{ as } E(u_i) = 0$$

Now, $Var(Y_i) = E[\alpha + \beta X_i + u_i - \alpha - \beta X_i]^2$

$$\left. = E(u_i^2) = \sigma_u^2 \right]$$

Similarly, we may obtain

$$\beta^* = \sum_{i=1}^{n} C_i Y_i \text{ and } Var(\beta^*) = Var\left(\sum_{i=1}^{n} C_i Y_i\right) = \sum_{i=1}^{n} C_i^2 Var(Y_i) = \sum_{i=1}^{n} C_i^2 \sigma_u^2$$

Now, $\sum_{i=1}^{n} C_i^2 = \sum_{i=1}^{n} (K_i + d_i)^2$

$$= \sum_{i=1}^{n} K_i^2 + \sum_{i=1}^{n} d_i^2 + 2\sum_{i=1}^{n} K_i d_i = \sum_{i=1}^{n} K_i^2 + \sum_{i=1}^{n} d_i^2$$

Given that $\sum_{i=1}^{n} K_i d_i = \dfrac{\sum_{i=1}^{n} x_i d_i}{\sum_{i=1}^{n} x_i^2} = \dfrac{\sum_{i=1}^{n} (X_i - \bar{X}) d_i}{\sum_{i=1}^{n} x_i^2}$

$$= \frac{\sum_{i=1}^{n} d_i X_i - \bar{X}\sum_{i=1}^{n} d_i}{\sum_{i=1}^{n} x_i^2} = 0 \text{ (as } \sum_{i=1}^{n} d_i X_i = 0 \text{ and } \sum_{i=1}^{n} d_i = 0)$$

Substituting we find

$$Var(\beta^*) = \sigma_u^2\left(\sum_{i=1}^{n} K_i^2 + \sum_{i=1}^{n} d_i^2\right) = \sigma_u^2 \sum_{i=1}^{n} K_i^2 + \sigma_u^2 \sum_{i=1}^{n} d_i^2$$

$$\therefore Var(\beta^*) = Var(\hat{\beta}) + \sigma_u^2 \sum_{i=1}^{n} d_i^2 \left[\because Var(\hat{\beta}) = \sigma_u^2 \sum_{i=1}^{n} K_i^2 = \sigma_u^2 / \Sigma x_i^2\right]$$

Since $\sigma_u^2\sum_{i=1}^{n} d_i^2 > 0$, it proves that $Var(\beta^*) > Var(\hat{\beta})$ or, $Var(\hat{\beta}) < Var(\beta^*)$

Thus it is proved that $\hat{\beta}$ is the BLUE of β.

Case (b) In the same way it can be proved that the least squares constant intercept $\hat{\alpha}$ possesses minimum variance. We take a new estimator α^* which we assume to be a linear function of the Y_i's with weights $C_i = K_i + d_i$.

where $K_i = \dfrac{x_i}{\sum_{i=1}^{n} x_i^2}$.

Since $\hat{\alpha} = \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{X}K_i\right)Y_i$

Similarly, $\alpha^* = \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{X}C_i\right)Y_i = f(Y)$

This shows that like $\hat{\alpha}$, α^* is also a linear function in Y_i's.
Now α^* is to be regarded as an unbiased estimator of α if $E(\alpha^*) = \alpha$.
We substitute for $Y_i = \alpha + \beta X_i + u_i$ in α^* and we get,

$$\alpha^* = \alpha\left[1 - \bar{X}\sum_{i=1}^{n} C_i\right] + \beta\left[\bar{X} - \bar{X}\sum_{i=1}^{n} C_i X_i\right] + \sum_{i=1}^{n}\left[\frac{1}{n} - \bar{X}C_i\right]u_i$$

Now $E(\alpha^*) = \alpha\left[1 - \bar{X}E\left(\sum_{i=1}^{n} C_i\right)\right] + \beta\left[\bar{X} - \bar{X}E\left(\sum_{i=1}^{n} C_i X_i\right)\right] + E\left[\sum_{i=1}^{n}\left(\frac{1}{n} - \bar{X}C_i\right)u_i\right]$

Now $E(\alpha^*) = \alpha$ if and only if $\sum_{i=1}^{n} C_i = 0$, $\sum_{i=1}^{n} C_i X_i = 1$ and $\sum_{i=1}^{n} C_i u_i = 0$

These conditions imply $\sum_{i=1}^{n} d_i = 0$ and $\sum_{i=1}^{n} d_i X_i = 0$.

The variance of α^* is given by,
$Var(\alpha^*) = E[\alpha^* - E(\alpha^*)]^2 = E[\alpha^* - \alpha]^2$

$$= \sigma_u^2\sum_{i=1}^{n}\left[\frac{1}{n} - \bar{X}C_i\right]^2 = \sigma_u^2\sum_{i=1}^{n}\left[\frac{1}{n^2} - 2\frac{1}{n}\bar{X}C_i + \bar{X}^2C_i^2\right]$$

$$= \sigma_u^2\left[\frac{n}{n^2} - 2\bar{X}\frac{1}{n}\sum_{i=1}^{n} C_i + \bar{X}^2\sum_{i=1}^{n} C_i^2\right] = \sigma_u^2\left[\frac{1}{n} + \bar{X}^2\sum_{i=1}^{n} C_i^2 - \frac{2}{n}\cdot\bar{X}\sum_{i=1}^{n} C_i\right]$$

Since $\sum_{i=1}^{n} C_i = 0$ and $\sum_{i=1}^{n} C_i^2 = \sum_{i=1}^{n} K_i^2 + \sum_{i=1}^{n} d_i^2$

we have, $Var(\alpha^*) = \sigma_u^2\left[\frac{1}{n} + \overline{X}^2\left(\sum_{i=1}^{n} K_i^2 + \sum_{i=1}^{n} d_i^2\right)\right]$

$$= \sigma_u^2\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} x_i^2}\right] + \left[\sigma_u^2 \overline{X}^2 \sum_{i=1}^{n} d_i^2\right] \text{ where } \sum_{i=1}^{n} K_i^2 = \frac{1}{\sum_{j=1}^{n} x_i^2}$$

$$\therefore Var(\alpha^*) = \sigma_u^2\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} x_i^2}\right] + \left[\sigma_u^2 \overline{X}^2 \sum_{i=1}^{n} d_i^2\right] \text{ or, } Var(\alpha^*) = Var(\hat{\alpha}) + \sigma_u^2\left(\overline{X}^2 \sum_{i=1}^{n} d_i^2\right)$$

Here $\sum_{i=1}^{n} d_i^2 > 0$, because all d_i's are not zero.

Thus we have, $Var(\alpha^*) > Var(\hat{\alpha})$ or, $Var(\hat{\alpha}) < Var(\alpha^*)$.

Hence it is proved that $\hat{\alpha}$ is the BLUE of α.

2.10. The Variance of the Random Variable, u

The formulae of the variance of $\hat{\alpha}$ and $\hat{\beta}$ involve the variance of the random term u, σ_u^2. However, the true variance of u_i cannot be computed since the values of u_i are not observable. But we may obtain an unbiased estimate of σ_u^2 from the expression

$\hat{\sigma}_u^2 = \sum_{i=0}^{n} e_i^2 \Big/ (n-2)$ where $e_i = Y_i - \hat{Y}_i = Y_i - \hat{\alpha} - \hat{\beta}X_i$

[Y_i is the observed value and $\hat{Y}_i$ is the estimated value i.e., $Y_i = \alpha + \beta X_i + e_i$ and $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$ for $i = 1, 2, ..., n$]

Proof : One property of the regression line $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$ is that it passes through the point $(\overline{X}, \overline{Y})$. So, $\overline{Y} = \hat{\alpha} + \hat{\beta}\overline{X}$.

Again we know that $\overline{Y} = \alpha + \beta\overline{X} + \overline{u}$ from the observed relationship.

[Where $Y_i = \alpha + \beta X_i + u_i$ $\therefore \sum_{i=1}^{n} Y_i = n\alpha + \beta\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} u_i$

or, $\sum_{i=1}^{n} Y_i \Big/ n = \alpha + \beta\sum_{i=1}^{n} X_i \Big/ n + \sum_{i=1}^{n} u_i \Big/ n$ or, $\overline{Y} = \alpha + \beta\overline{X} + \overline{u}$]

Since $e_i = Y_i - \hat{Y}_i$

$$= (Y_i - \bar{Y}) - (\hat{Y}_i - \bar{Y}) = (\alpha + \beta X_i + u_i - \alpha - \beta \bar{X} - \bar{u}) - (\hat{\alpha} + \hat{\beta} X_i - \hat{\alpha} - \hat{\beta}\bar{X})$$

$$= [\beta(X_i - \bar{X}) + (u_i - \bar{u})] - [\hat{\beta}(X_i - \bar{X})]$$

$\therefore e_i = -(\hat{\beta} - \beta)x_i + (u_i - \bar{u})$ where $x_i = X_i - \bar{X}$

or, $e_i^2 = (\hat{\beta} - \beta)^2 x_i^2 + (u_i - \bar{u})^2 - 2x_i(\hat{\beta} - \beta)(u_i - \bar{u})$

$$= (\hat{\beta} - \beta)^2 x_i^2 + (u_i - \bar{u})^2 - 2(\hat{\beta} - \beta)(x_i u_i - x_i \bar{u})$$

$$\therefore \sum_{i=1}^{n} e_i^2 = (\hat{\beta} - \beta)^2 \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} (u_i - \bar{u})^2 - 2(\hat{\beta} - \beta)\left(\sum_{i=1}^{n} x_i u_i - \bar{u}\sum_{i=1}^{n} x_i\right)$$

$$= \left(\frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2}\right)^2 \cdot \sum_{i=1}^{n} x_i^2 + \left\{\sum_{i=1}^{n} u_i^2 - \left(\frac{\sum_{i=1}^{n} u_i}{n}\right)^2\right\} - 2\frac{\left(\sum_{i=1}^{n} x_i u_i\right)^2}{\sum_{i=1}^{n} x_i^2}$$

$\Bigg[$Since $\hat{\beta} = \beta + \dfrac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2}$ $\therefore\ \hat{\beta} - \beta = \sum_{i=1}^{n} x_i u_i \Big/ \sum_{i=1}^{n} x_i^2$

Again, $\sum_{i=1}^{n} (u_i - \bar{u})^2 = \sum_{i=1}^{n} u_i^2 - 2\bar{u}\sum_{i=1}^{n} u_i + \sum_{i=1}^{n} \bar{u}^2$

$$= \sum_{i=1}^{n} u_i^2 - 2\bar{u} \cdot n \frac{1}{n}\sum_{i=1}^{n} u_i + n\bar{u}^2 = \sum_{i=1}^{n} u_i^2 - 2n\bar{u}^2 + n\bar{u}^2$$

$$= \sum_{i=1}^{n} u_i^2 - n\bar{u}^2 = \sum_{i=1}^{n} u_i^2 - n\left(\sum_{i=1}^{n} \frac{u_i}{n}\right)^2 = \sum_{i=1}^{n} u_i^2 - \frac{\left(\sum u_i^2\right)^2}{n}$$

and $2(\hat{\beta} - \beta)\left(\sum_{i=1}^{n} x_i u_i - \bar{u}\sum_{i=1}^{n} x_i\right) = 2\dfrac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2} \cdot \sum_{i=1}^{n} x_i u_i = 2\left(\sum_{i=1}^{n} x_i u_i\right)^2 \Big/ \sum_{i=1}^{n} x_i^2 \Bigg]$ as $\bar{u} = 0$.

$$\therefore \sum_{i=1}^{n} e_i^2 = \left\{\sum_{i=1}^{n} u_i^2 - \frac{\left(\sum_{i=1}^{n} u_i\right)^2}{n}\right\} - \left(\sum_{i=1}^{n} x_i u_i\right)^2 \Big/ \sum_{i=1}^{n} x_i^2$$

$$\text{or, } \sum_{i=1}^{n} e_i^2 = \left[\sum_{i=1}^{n} u_i^2 - \frac{\left(\sum_{i=1}^{n} u_i^2 + 2\sum_i\sum_j u_i u_j\right)}{n}\right] - \left[\frac{\sum_{i=1}^{n} x_i^2 u_i^2 + 2\sum_i\sum_j x_i x_j u_i u_j}{\sum_{i=1}^{n} x_i^2}\right]$$

$$\therefore E\left[\sum_{i=1}^{n} e_i^2\right] = \left[\sum_{i=1}^{n} E(u_i^2) - \frac{\left\{\sum_{i=1}^{n} E(u_i^2) + 2\sum_i\sum_j E(u_i u_j)\right\}}{n}\right] - \left[\frac{\sum_{i=1}^{n} x_i^2 E(u_i^2) + 2\sum_i\sum_j x_i x_j E(u_i u_j)}{\sum_{i=1}^{n} x_i^2}\right]$$

$$= \left[\sum_{i=1}^{n} \sigma_u^2 - \sum_{i=1}^{n} \frac{\sigma_u^2}{n}\right] - \left[\sum_{i=1}^{n} x_i^2 \sigma_u^2\right] \Big/ \sum_{i=1}^{n} x_i^2 \quad \left[\because E(u_i^2) = \sigma_u^2 \text{ and } E(u_i u_j) = 0\right]$$

$$\text{or, } E\left(\sum_{i=1}^{n} e_i^2\right) = \left[n\sigma_u^2 - \frac{n\sigma_u^2}{n}\right] - \sigma_u^2 \sum_{i=1}^{n} x_i^2 \Big/ \sum_{i=1}^{n} x_i^2$$

$$= n\sigma_u^2 - \sigma_u^2 - \sigma_u^2 = n\sigma_u^2 - 2\sigma_u^2 = \sigma_u^2(n-2)$$

$$\text{or, } E\left(\frac{\sum_{i=1}^{n} e_i^2}{n-2}\right) = \sigma_u^2$$

So, $\sum_{i=1}^{n} e_i^2/(n-2)$ is an unbiased estimator of σ_u^2. If we denote $\sum_{i=1}^{n} e_i^2/(n-2) = \hat{\sigma}_u^2$ then $\hat{\sigma}_u^2$ is an unbiased estimator of σ_u^2.

2.11. Maximum Likelihood Estimators (MLE's) of α, β and σ_u^2

If each $u_i [Y_i = \alpha + \beta X_i + u_i]$ is normally distributed with mean 0 and variance σ_u^2 i.e., $u_i \sim N(0, \sigma_u^2)$ and $u_1, u_2, \ldots u_n$ are independent, then MLE of α and β are equivalent to the OLS estimators of α and β (i.e., $\hat{\alpha}$ and $\hat{\beta}$).

Proof : Since $u_i \sim N(0, \sigma_u^2)$, the p.d.f. of u_i is given by,

$$f_i(u_i) = \frac{1}{\sqrt{2\pi}\sigma_u} e^{-\frac{1}{2}\left(\frac{u_i - \bar{u}}{\sigma_u}\right)^2} = \frac{1}{\sqrt{2\pi}\sigma_u} \cdot e^{-\frac{1}{2\sigma_u^2}u_i^2}, \text{ as } \bar{u} = 0.$$

This joint probability distribution function of $u_1, u_2, \ldots, u_n$ is given by $f(u_1, u_2, \ldots u_n)$ and given the set of sample observations it is looked upon as a function of the parameters and is called the likelihood function of the parameters. Since $u_1, u_2, \ldots, u_n$ are independent, then we can write,

$$f(u_1, u_2, \ldots, u_n) = \underset{i=1}{\overset{n}{\pi}} f_i(u_i)$$

$$\text{or, } L(\alpha, \beta, \sigma_u^2) = \underset{i=1}{\overset{n}{\pi}} f_i(u_i) \text{ or, } L(\alpha, \beta, \sigma_u^2) = \underset{i=1}{\overset{n}{\pi}} \frac{1}{\sqrt{2\pi\sigma_u}} e^{-\frac{1}{2\sigma_u^2}u_i^2}$$

$$\text{or, } L(\alpha, \beta, \sigma_u^2) = \frac{1}{(\sqrt{2\pi})^n \sigma_u^n} \cdot e^{-\frac{1}{2\sigma_u^2}\sum_{i=1}^{n} u_i^2}$$

Taking Log on both sides we get,

$$\log L = -\frac{n}{2}\log 2\pi - n\log\sigma_u - \frac{1}{2\sigma_u^2} \cdot \sum_{i=1}^{n} u_i^2$$

$$= -\frac{n}{2}\log 2\pi - n\log\sigma_u - \frac{1}{2\sigma_u^2} \cdot \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2$$

$$[\because Y_i = \alpha + \beta X_i + u_i \text{ or, } u_i = Y_i - \alpha - \beta X_i \text{ or, } \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2]$$

MLE of α and β can be obtained by maximising log L through the choice of α and β. Maximisation of log L through the choice of α and β is equivalent to minimisation of $\sum_{i=1}^{n}(Y_i - \alpha - \beta X_i)^2$ through the choice of α and β.

Let us suppose that α^* is the MLE of α and β^* is the MLE of β. Then,

$$\beta^* = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}, \alpha^* = \overline{Y} - \beta^* \overline{X} \qquad \left[\text{Since } \hat{\beta} = \sum_{i=1}^{n} x_i y_i \Big/ \sum_{i=1}^{n} x_i^2 \text{ and } \hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}\right]$$

$$\text{Since } \log L = -\frac{n}{2}\log 2\pi - n\log\sigma_u - \frac{1}{2\sigma_u^2}\sum_{i=1}^{n}(Y_i - \alpha - \beta X_i)^2$$

Differentiating partially log L with respect to α, β we get,

$$\frac{\delta \log L}{\delta \alpha} = -\frac{1}{2\sigma_u^2} 2\sum_{i=1}^{n}(Y_i - \alpha - \beta X_i)(-1)$$

$$\frac{\delta \log L}{\delta \log \beta} = -\frac{1}{2\sigma_u^2}\sum_{i=1}^{n}(Y_i - \alpha - \beta X_i) \cdot (-X_i)$$

Equating these equations to zero and putting 'star' mark on the parameters to distinguish them from least squares estimators,

we get,

$$\frac{1}{2\sigma_u^2}\sum_{i=1}^{n}(Y_i-\alpha^*-\beta^*X_i)=0 \qquad \text{... (1)}$$

$$\frac{1}{2\sigma_u^2}\sum_{i=1}^{n}X_i(Y_i-\alpha^*-\beta^*X_i)=0 \qquad \text{... (2)}$$

The first two equations are reduced to the least squares normal equations :

$$\sum_{i=1}^{n}Y_i=n\alpha^*+\beta^*\sum_{i=1}^{n}X_i$$

$$\sum_{i=1}^{n}X_iY_i=\alpha^*\sum_{i=1}^{n}X_i+\beta^*\sum_{i=1}^{n}X_i^2$$

Now solving the two normal equations we can get $\beta^*\frac{\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}x_i^2}$ and $\alpha^*=\bar{Y}-\beta^*\bar{X}$

This proves that the MLE of α and β are the same as the least squares estimators. Hence, they would also possess all the desirable properties.

When $\log L$ is maximised through the choice of α and β and α^* and β^* are the MLEs of α and β, then,

$$\sum_{i=1}^{n}u_i^2=\sum_{i=1}^{n}(Y_i-\alpha^*-\beta^*X_i)^2 \qquad [\because \alpha^*=\hat{\alpha} \text{ and } \beta^*=\hat{\beta}]$$

Hence, $\sum_{i=1}^{n}u_i^2=\sum_{i=1}^{n}\left(Y_i-\hat{\alpha}-\hat{\beta}X_i\right)^2=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2=\sum_{i=1}^{n}e_i^2$ $\qquad [\therefore \hat{Y}_i=\hat{\alpha}+\hat{\beta}X_i]$

$$\therefore \sum_{i=1}^{n}u_i^2=\sum_{i=1}^{n}e_i^2$$

So, the likelihood function maximised with respect to α and β is given by $\log L=-\frac{n}{2}\log 2\pi-n\log\sigma_u-\frac{1}{2\sigma_u^2}\cdot\sum_{i=1}^{n}e_i^2$.

In order to obtain the MLE of σ_u^2, $\log L$ is to be maximised through the choice of σ_u and the first order condition of maximisation is given by (assuming σ_u^2* as the MLE of σ_u^2).

$$\frac{\delta\log L}{\delta\sigma_u}=-\frac{n}{\sigma_u}-\frac{1}{2}\sum_{i=1}^{n}e_i^2(-2)\cdot\frac{1}{\sigma_u^3}=0=\frac{1}{\sigma_u}\left[-n+\left(\sum_{i=1}^{n}e_i^2\Big/\sigma_u^2\right)\right]=0$$

or, $\sum_{i=1}^{n}e_i^2\Big/\sigma_u^2=n$ or, $\sigma_u^2=\sum_{i=1}^{n}e_i^2\Big/n=\sigma_u^2*$ (say) so, $\sum_{i=1}^{n} e_i^2 \Big/ n = \sigma_u^2 *$, (say), is the MLE of the variance of the disturbance term, denoted by $\sigma_u^2 *$.

[*Note* : For maximisation, however, we require second order conditions. But this is not shown here. But we should also check the second order conditions, required for maximisation. This means that we have to show that, $\frac{\delta^2 \log L}{\delta \alpha^2} < 0$, $\frac{\delta^2 \log L}{\delta \beta^2} < 0$ and $\frac{\delta^2 \log L}{\delta \sigma_u^2} < 0$.]

Thus we see that the MLE of σ_u^2 i.e., $\sigma_u^2 * = \sum_{i=1}^{n} e_i^2 \Big/ n$ is not an unbiased estimator but it is a consistent estimator.

$$\text{i.e.,}\quad E\left(\sum_{i=1}^{n} e_i^2 \Big/ n\right) = E\left[\left(\sum_{i=1}^{n} e_i^2 \Big/ (n-2)\right) \times \frac{n-2}{n}\right]$$

$$= E\left[\left(\sum_{i=1}^{n} e_i^2 \Big/ (n-2)\right)\left(\frac{n-2}{n}\right)\right] = E\left[\left(\sum_{i=1}^{n} e_i^2 \Big/ (n-2)\right)\left(1 - \frac{2}{n}\right)\right]$$

$$= \sigma_u^2\left[1 - \frac{2}{n}\right], \text{ since } E\left(\sum_{i=1}^{n} e_i^2 \Big/ (n-2)\right) = \sigma_u^2 .$$

$$\therefore E\left(\sum_{i=1}^{n} e_i^2 \Big/ n\right) = \sigma_u^2\left(1 - \frac{2}{n}\right)$$

$$\text{Now } E\left[\sum_{i=1}^{n} e_i^2 \Big/ n\right] \to \sigma_u^2 \text{ as } n \to \infty.$$

This proves that the MLE of σ_u^2 i.e., $\sum_{i=1}^{n} e_i^2 \Big/ n$ is a consistent estimator of σ_u^2.

Note : MLE of α and β i.e., $\alpha*$ and $\beta*$ are unbiased estimators of α and β but MLE of σ_u^2 i.e., $\sigma_u^2 * = \sum_{i=1}^{n} e_i^2 \Big/ n$ is not an unbiased estimator, rather it is a consistent estimator (consistently unbiased) of σ_u^2.

2.12. The Sampling Distribution of the Least Squares Estimates

Since least squares estimators are linear combinations of independent normal variables, $Y_1, Y_2, ..., Y_n$; $\hat{\alpha}$ and $\hat{\beta}$ must also be normally distributed with the following characteristics :

(i) $\hat{\alpha}$ and $\hat{\beta}$ are unbiased estimators, their means being equal to their true values of α and β.

(ii) Variance of each estimator is known.

Both these results may be stated symbolically as follows :

$$\hat{\alpha} \sim N\left[\alpha, \sigma_u^2\left(\frac{1}{n}+\frac{\overline{X}^2}{\sum_{i=1}^{n} x_i^2}\right)\right] \Rightarrow E(\hat{\alpha}) = \alpha,\ \text{var}(\hat{\alpha}) = \sigma_u^2\left(\frac{1}{n}+\frac{\overline{X}^2}{\sum_{i=1}^{n} x_i^2}\right)$$

$$\hat{\beta} \sim N\left[\beta, \frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2}\right] \Rightarrow E(\hat{\beta}) = \beta,\ \text{var}(\hat{\beta}) = \frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2}.$$

Variances of the parameters are directly related to the variances of the disturbances. The following points should be noted carefully :

(i) Larger the value of σ_u^2, the larger the variances of $\hat{\alpha}$ and $\hat{\beta}$. In other words, the greater the dispersion of the disturbance terms around the population regression line, the greater is the dispersion in the values of estimated regression parameters.

(ii) $\sum_{i=1}^{n} x_i^2$ is in the denominator of the variance formula of both the estimators. This indicates that the more dispersed the values of the explanatory variables (i.e., larger $\sum_{i=1}^{n} x_i^2$), the smaller the variances of $\hat{\alpha}$ and $\hat{\beta}$. If $\sum_{i=1}^{n} x_i^2$ tends to zero ; i.e., when $X_1 = X_2 = ... = X_n$, both variances would be infinitely large.

(iii) The variance of $\hat{\alpha}$ is the smallest when $\overline{X} = 0$ or tends to zero. In particular, when $\overline{X} = 0$, var $(\hat{\alpha}) = \frac{\sigma_u^2}{n}$.

2.13. Confidence Intervals and Hypothesis Testing

It is highly essential to construct confidence intervals of the parameters in order to achieve precision of $\hat{\alpha}$ and $\hat{\beta}$. We have all the information concerning the distribution of $\hat{\alpha}$ and $\hat{\beta}$ in order to standardise them.

Since $\hat{\alpha} \sim N\left[\alpha, \sigma_u^2\left(\frac{1}{n}+\frac{\overline{X}^2}{\sum_{i=1}^{n} x_i^2}\right)\right]$ and $\hat{\beta} \sim N\left(\beta, \frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2}\right)$

Now τ or $Z = \dfrac{\hat{\beta} - E(\hat{\beta})}{SE(\hat{\beta})} = \dfrac{\hat{\beta} - \beta}{\sigma_u \sqrt{\dfrac{1}{\sum_{i=1}^{n} x_i^2}}} \sim N\,(0,\,1),$

where $SE(\hat{\beta}) = \sqrt{\text{var}(\hat{\beta})}$ and τ or $Z = \dfrac{\hat{\alpha} - E(\hat{\alpha})}{SE(\hat{\alpha})} = \dfrac{\hat{\alpha} - \alpha}{\sigma_u \sqrt{\dfrac{\sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} x_i^2}}} \sim N(0,1).$

Here $\text{var}(\hat{\alpha}) = \sigma_u^2 \left(\dfrac{1}{n} + \dfrac{\bar{X}^2}{\sum_{i=1}^{n} x_i^2} \right).$

$$\therefore SE(\hat{\alpha}) = \sqrt{\text{var}(\hat{\alpha})} = \sigma_u \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} x_i^2}} = \sigma_u \sqrt{\sum_{i=1}^{n} X_i^2 \Big/ n \sum_{i=1}^{n} x_i^2}$$

$$\left[\because \frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i^2 + n\bar{X}^2}{n \sum_{i=1}^{n} x_i^2} \right.$$

$$= \frac{\left[\sum_{i=1}^{n} (X_i - \bar{X})^2 \right] + n\bar{X}^2}{n \sum_{i=1}^{n} x_i^2} = \frac{\left[\sum_{i=1}^{n} X_i^2 - 2\bar{X} \sum_{i=1}^{n} X_i + n\bar{X}^2 \right] + n\bar{X}^2}{n \sum_{i=1}^{n} x_i^2}$$

$$\left. = \frac{\sum_{i=1}^{n} X_i^2 - 2n\bar{X}^2 + 2n\bar{X}^2}{n \sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} x_i^2} \right]$$

Here σ_u^2 represents the variance of the unobservable disturbances which is known. In particular, if σ_u^2 is not known and we substitute by its unbiased estimator $\sum_{i=1}^{n} e_i^2 \Big/ (n-2)$ $\left[i.e., E(\sigma_u^2) = \sum_{i=1}^{n} e_i^2 \Big/ (n-2) = \hat{\sigma}_u^2 \right]$, then the standard normal variable Z or τ will follow a *t*-distribution with $(n-2)$ degrees of freedom.

In case of $\hat{\alpha}$, $\tau = \dfrac{\hat{\alpha} - \alpha}{\sigma_u \sqrt{\dfrac{\sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} x_i^2}}} = \dfrac{(\hat{\alpha} - \alpha)\sqrt{n \sum_{i=1}^{n} x_i^2}}{\sigma_u \sqrt{\sum_{i=1}^{n} X_i^2}}$.

When σ_u is not known and it is replaced by $\hat{\sigma}_u$ $\left[\hat{\sigma}_u = \sqrt{\sum_{i=1}^{n} e_i^2 / (n-2)} \right]$, the unbiased estimator of σ_u, then we have,

$$t = t_{n-2} = \frac{(\hat{\alpha} - \alpha)\sqrt{n \sum_{i=1}^{n} x_i^2}}{\hat{\sigma}_u \sqrt{\sum_{i=1}^{n} X_i^2}} \text{ with d.f} = (n-2).$$

Now by rearranging in terms of '*t*' expression we have :

$$\hat{\alpha} - \alpha = t_{n-2} \hat{\sigma}_u \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} x_i^2}} \quad \text{or,} \quad \alpha = \hat{\alpha} \pm t_{n-2} \hat{\sigma}_u \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} x_i^2}}$$

Therefore 95% confidence limits for α are :

$$\hat{\alpha} \pm t_{0.025_{n-2}} \cdot \hat{\sigma}_u \sqrt{\sum_{i=1}^{n} X_i^2 \Big/ n \sum_{i=1}^{n} x_i^2}$$

Similarly, 99% confidence limits for α are :

$$\hat{\alpha} \pm t_{0.005_{n-2}} \cdot \hat{\sigma}_u \sqrt{\sum_{i=1}^{n} X_i^2 \Big/ n \sum_{i=1}^{n} x_i^2}$$

[The values of $t_{0.025_{n-2}}$ and $t_{0.005_{n-2}}$ corresponding to $(n-2)$ d.f can be obtained from the table, given at the end of the book.]

In the same way, for testing β, we have,

$$t = \frac{\hat{\beta}-\beta}{\sigma_u\sqrt{\dfrac{1}{\sum_{i=1}^{n} x_i^2}}} = \frac{(\hat{\beta}-\beta)\sqrt{\sum_{i=1}^{n} x_i^2}}{\sigma_u}$$

when σ_u is not known then it is replaced by its unbiased estimator $\hat{\sigma}_u$, then we have :

$$t = t_{n-2} = \frac{(\hat{\beta}-\beta)\sqrt{\sum_{i=1}^{n} x_i^2}}{\hat{\sigma}_u}$$ with d.f = $n-2$. Now rearranging we may get,

$$\hat{\beta}-\beta = t_{n-2}\cdot\frac{\hat{\sigma}_u}{\sqrt{\sum_{i=1}^{n} x_i^2}} \quad \text{or,} \quad \beta = \hat{\beta} \pm t_{n-2}\cdot\frac{\hat{\sigma}_u}{\sqrt{\sum_{i=1}^{n} x_i^2}}$$

Therefore 95% confidence limits for β would be :

$$\hat{\beta} \pm t_{0.025,n-2}\cdot\frac{\hat{\sigma}_u}{\sqrt{\sum_{i=1}^{n} x_i^2}}$$

and 99% confidence limits for β would be :

$$\hat{\beta} \pm t_{0.005,n-2}\cdot\frac{\hat{\sigma}_u}{\sqrt{\sum_{i=1}^{n} x_i^2}}$$

Confidence interval for σ_u^2 :

Under the normality assumption, the variable $\chi^2 = (n-2)\dfrac{\hat{\sigma}_u^2}{\sigma_u^2}$ follows a χ^2 distribution with d.f = $(n-2)$.

Therefore, we can use χ^2_{n-2} to establish a confidence interval for σ_u^2,

$$P\left[\chi^2_{1-\frac{\alpha}{2}} \le \chi^2_{n-2} \le \chi^2_{\alpha/2}\right] = 1-\alpha$$

or, $$P\left[\chi^2_{1-\frac{\alpha}{2}} \le (n-2)\frac{\hat{\sigma}_u^2}{\sigma_u^2} \le \chi^2_{\alpha/2}\right] = 1-\alpha$$

or, $$P\left[(n-2)\frac{\hat{\sigma}_u^2}{\chi^2_{\alpha/2}} \le \sigma_u^2 \le (n-2)\frac{\hat{\sigma}_u^2}{\chi^2_{1-\alpha/2}}\right] = 1-\alpha$$

which gives 100 (1 – α)% confidence interval for σ_u^2.

So, 100 (1 – α)% confidence limits to σ_u^2 would be :

$$(n-2)\frac{\hat{\sigma}_u^2}{\chi^2_{\alpha/2}} \text{ and } (n-2)\frac{\hat{\sigma}_u^2}{\chi^2_{1-\alpha/2}}, \text{ usually } \alpha = 0.01 \text{ or, } 0.05.$$

All χ^2 values are taken from table with d.f = $(n-2)$.

Usual test procedure : One hypothesis of usual interest is that we hypothesise that there is no relationship (linear) between the explanatory variable X and the dependent variable Y in the regression model : $Y = \alpha + \beta X + u$.

As such the null hypothesis of no relationship between X and Y is $H_0 : \beta = 0$ and we have to test it against the alternative hypothesis $H_1 : \beta \ne 0$.

Now the appropriate test statistic under $H_0 : \beta = 0$ would be

$$t = t_{n-2} = \frac{\hat{\beta}-0}{SE(\hat{\beta})} = \frac{\hat{\beta}}{\hat{\sigma}_u \Big/ \sqrt{\sum_{i=1}^{n} x_i^2}} = \frac{\hat{\beta}\sqrt{\sum_{i=1}^{n} x_i^2}}{\hat{\sigma}_u}$$

which follows a t-distribution with d.f = $(n-2)$.

Now at 5% level of significance the null hypothesis will be rejected for the given sample if $|t_{n-2}|$ (observed) $> t_{0.025,\ n-2}$ and will be accepted otherwise (i.e., if $-t_{0.025,n-2} \le t \le t_{0.025,n-2}$). Similarly, at 1% level of significance the null hypothesis will be rejected for the given sample if $|\ t_{n-2}|$ (observed) $> t_{0.005,\ n-2}$ and will be accepted otherwise (i.e., if $-t_{0.005,n-2} \le t \le t_{0.005,n-2}$). The confidence limits for β (acceptance region in a two tailed test) at 5% and 1% levels of significance with $(n-2)$ degrees of freedom will be given by,

$$-t_{0.025,n-2}.SE(\hat{\beta}) \le \beta \le +t_{0.025,n-2}.SE(\hat{\beta})$$

and $$-t_{0.005,n-2}.SE(\hat{\beta}) \le \beta \le +t_{0.005,n-2}.SE(\hat{\beta})$$

where $SE(\hat{\beta}) = \dfrac{\hat{\sigma}_u}{\sqrt{\sum_{i=1}^{n} x_i^2}}$ (σ_u is not known, and replaced by $\hat{\sigma}_u$).

2.13.1. The Exact Level of Significance : The *p*-value

We know that the significance level α in a hypothesis testing problem is the probability of making a Type-I error, i.e., α is the probability of rejecting the null hypothesis H_0 : when it is true. Traditionally, we fix the value of α as the maximum allowable value for the probability of Type-I error and then proceed to test the null hypothesis in terms of the appropriate test statistic (τ, t, χ^2 etc). The pre specified value of α (usually $\alpha = 0.01$, or 0.05 and 0.10 in some cases) sets corresponding to tolerance limits i.e., the critical value(s) for the observed test statistic. If the calculated value of the test statistic exceeds the tolerance limit(s) set forth by the pre-specified α then H_0 is rejected and other wise accepted. Smaller the value of α, the more liberal the tolerance limit for the observed test statistic gets.

Now we can ask a natural question : what is the smallest value of α (significance level) at which the null hypotheis, gets rejected ? The answer is p-value (or probability value) associated with the observed data set.

Once the p-value is computed, one can make use of it to come to a conclusion by comparing the p-value with α.

Usually, if the p-value $\leq \alpha$ then reject H_0 ; otherwise accept H_0.

To illustrate it, we consider an example where we have fitted the regression equation $Y_i = \alpha + \beta X_i + u_i$, showing the impact of education on wages, given a sample size of $n = 13$ where Y = wages, X = Education/years of schooling (the original data are not given here).

The estimated regression results are given below :

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

$$\Rightarrow \quad \hat{Y}_i = -0.0144 + 0.7240 X_i$$

$$\text{S.E} : (0.9317) \; (0.0700), \; r^2 = 0.9065$$

Suppose we like to test the null hypothesis $H_0 : \beta = 0.5$ against the alternative hypothesis $H_1 : \beta \neq 0.5$. The appropriate test statistic under $H_0 : \beta = 0.5$ would be

$$t_{n-2} = \frac{\hat{\beta}-\beta}{SE(\hat{\beta})} = \frac{0.7240-0.5}{0.0700} = 3.2$$

$\therefore$ t_{n-2} (observed) = 3.2

Now on the basis of the given sample $H_0 : \beta = 0.5$ will be rejected at 100 α% level of significance when ($H_1 : \beta \neq 0.5$) if | t_{n-2} (observed)| > $t_{\alpha/2,n-2}$ (Table)

Here $n = 13$, if $\alpha = 0.05$, $t_{\alpha/2,n-2} = t_{0.025,11} = 2.201$

and if $\alpha = 0.01$, $t_{\alpha/2,n-2} = t_{0.005,11} = 3.106$

So, $H_0 : \beta = 0.5$ is rejected both at 5% and 1% levels of significane as | t_{n-2} (observed)| = 3.2 > 2.201 and 3.106.

Now on the basis of p-value the null hypothesis will be rejected for the given sample if $p \leq \alpha$ and will be accepted otherwise.

Here given the null hypothesis, that the true coefficient of education ($\beta = 0.5$), we obtain a t value of 3.2. Now what is the p-value of obtaining a t value of as much as or greater than 3.2 ?

From the 't' table given in Appendix Table-IV we see that for 11 d.f. the probability of obtaining such 't' value must be smaller than 0.005 (one tail) or 0.010 (two tail).

If we use Stata or *E*-views statistical packages, we will find that the *p*-value of obtaining a '*t*' value of 3.2 or greater is about 0.00001, or $P\,[\,|\,t\,|> 3.2\,] = 0.00001$ that is, extremely small. This is the *p*-value of the observed '*t*' statistic. (i.e., Here $\alpha = 0.01$ and $p = 0.00001$ hence $p < \alpha$).

The exact level of significance of the '*t*' statistic is much smaller than the conventionally, and arbitrarily fixed level of significance, such as 1, 5 or 10 percent.

As a matter of fact, if we were to use the *p* value just computed, and reject the null hypothesis that the true coefficient of education is 0.5, the probability of committing a Type-I error would be only about 1 in 100,000 !

For a given sample size, as $|\,t\,|$ increases the *p*-value decreases, and we can therefore reject the null hypothesis with increasing confidence.

What is the relationship of the *p*-value to the level of significance α ? If we make the habit of fixing α equal to *p*-value of a test statistic (e.g., the '*t*' statistic) then there is no conflict between the two values. To put it differently it is better to give up fixing α arbitrarily at some level and simply choose the *p*-value of the test statistic. Usually we reject the null hypothesis, H_0 on the basis of the given sample if *p*-value $\leq \alpha$ and accept H_0 otherwise.

2.14. Goodness of Fit of the Multiple Correlation Coefficient (R^2)

So far we were concerned with the estimation and precision of the regression parameters α and β. We now like to consider the regression line as a whole and examine its goodness of fit. Suppose, a sample regression has been obtained by the method of least squares, as shown in the following diagram (Fig. 2.5). Considering a specific observation of the dependent variable Y_i, we can write : $e_i = Y_i - \hat{Y}_i$ where $Y_i = \alpha + \beta X_i + e_i$ and $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ (e_i being the error of estimate).

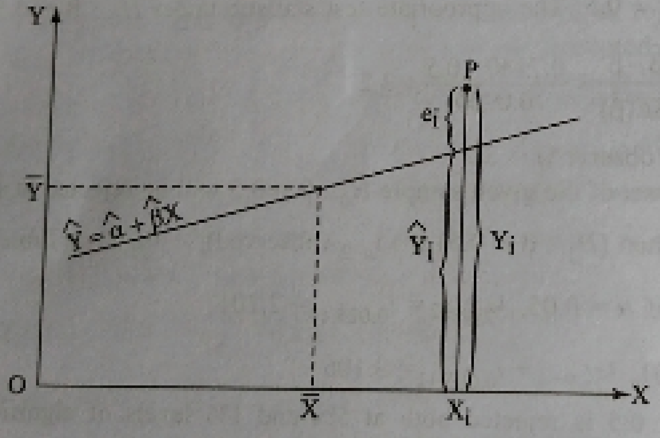


Fig. 2.5

Now $Y_i = \hat{Y}_i + e_i$

or, $Y_i - \overline{Y} = (\hat{Y}_i - \overline{Y}) + e_i - \overline{e}$ or, $(Y_i - \overline{Y}) = (\hat{Y}_i - \overline{Y}) + e_i$ $(\because \overline{e} = 0)$

or, $\sum_{i=1}^{n} (Y_i - \overline{Y}) = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y}) + \sum_{i=1}^{n} e_i$ and $\sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} \left[(\hat{Y}_i - \overline{Y}) + e_i \right]^2$

or, $\sum_{i=1}^{n}(Y_i-\bar{Y})^2=\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2+2\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})e_i+\sum_{i=1}^{n}e_i^2$

Now $\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})e_i=\sum_{i=1}^{n}(\hat{\alpha}+\hat{\beta}X_i-\bar{Y})e_i$

$$=\hat{\alpha}\sum_{i=1}^{n}e_i+\hat{\beta}\sum_{i=1}^{n}X_ie_i-\bar{Y}\sum_{i=1}^{n}e_i=0\left(\because\sum_{i=1}^{n}e_i=0\text{ and }\sum_{i=1}^{n}X_ie_i=0\right)$$

Again, $\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2=\sum_{i=1}^{n}(\hat{\alpha}+\hat{\beta}X_i-\bar{Y})^2$

$$=\sum_{i=1}^{n}\left[(\bar{Y}-\hat{\beta}\bar{X})+\hat{\beta}X_i-\bar{Y}\right]^2\quad[\because\hat{\alpha}=\bar{Y}-\hat{\beta}\bar{X}]$$

$$=\sum_{i=1}^{n}\left[\hat{\beta}(X_i-\bar{X})\right]^2=\hat{\beta}^2\sum_{i=1}^{n}(X_i-\bar{X})^2$$

$$\therefore\sum_{i=1}^{n}(Y_i-\bar{Y})^2=\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2+2\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})e_i+\sum_{i=1}^{n}e_i^2$$

$$=\hat{\beta}^2\sum_{i=1}^{n}(X_i-\bar{X})^2+2\times0+\sum_{i=1}^{n}e_i^2$$

or, $\sum_{i=1}^{n}(Y_i-\bar{Y})^2=\hat{\beta}^2\sum_{i=1}^{n}(X_i-\bar{X})^2+\sum_{i=1}^{n}e_i^2$

or, $\sum_{i=1}^{n}y_i^2=\hat{\beta}^2\sum_{i=1}^{n}x_i^2+\sum_{i=1}^{n}e_i^2$ where $y_i=Y_i-\bar{Y}$ and $x_i=X_i-\bar{X}$.

$$\Rightarrow\left\{\begin{matrix}\text{Total sum of}\\\text{squares}\end{matrix}\right\}=\left[\begin{matrix}\text{Explained sum}\\\text{of squares}\end{matrix}\right]+\left[\begin{matrix}\text{Unexplained}\\\text{sum of squares}\end{matrix}\text{ or, }\begin{matrix}\text{Residual sum}\\\text{of squares}\end{matrix}\right]$$

$\Rightarrow$ TSS = ESS + RSS

$\sum_{i=1}^{n}(Y_i-\bar{Y})^2$ represents the total sum of squared deviations from $\bar{Y}$, which we may take as a measure of the total variations in Y.

Thus total variations can be decomposed into two parts :

(i) $\hat{\beta}^2\left[\sum_{i=1}^{n}(X_i-\bar{X})^2\right]=\hat{\beta}^2\sum_{i=1}^{n}x_i^2\Rightarrow$ the estimated effect of change in X on the variations in Y (Explained sum of squares).

(ii) $\sum_{i=1}^{n} e_i^2 \Rightarrow$ the variations in Y which remain unexplained by the estimated relationship between Y and X (Unexplained sum of squares).

This decomposition of total variations in Y leads to a measure of the "goodness of fit"—which is also known as the coefficient of determination and is denoted by R^2,

$$\text{where } R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\text{Explained variatio}}{\text{Total variations}} \overset{\text{n}}{=} \frac{\text{sv}\hat{\text{a}}\text{r}}{\text{var}(Y)}$$

$$= \frac{\hat{\beta}^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2} = \frac{\sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} y_i^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} y_i^2} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

Since $\text{var}(Y) = \text{var}(\hat{Y}) + \text{var}(e)$

and $0 \leq \text{var}(\hat{Y}) \leq \text{var}(Y)$

or, $0 \leq \dfrac{\text{var}(\hat{Y})}{\text{var}(Y)} \leq 1$ or, $0 \leq R^2 \leq 1$

$R^2 = 0$ when $\text{var}(\hat{Y}) = 0$ i.e., $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} y_i^2$

and $R^2 = 1$ when $\text{var}(\hat{Y}) = \text{var}(Y)$ i.e., $\sum_{i=1}^{n} e_i^2 = 0$.

It should be noted that, this R^2 is equal to the square of the simple correlation coefficient between X and Y.

By definition simple correlation coefficient (product moment) is given by,

$$r_{XY} = r = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\frac{1}{n}\sum(Y_i - \bar{Y})^2}}$$

$$\therefore r = \frac{\sum_{k=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i^2}}$$

Since $\hat{\beta} = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$ and $R^2 = \dfrac{\hat{\beta}^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2}$

$$\therefore R^2 = \left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}\right)^2 \cdot \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2} = \frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2}{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2} = \left(\frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}\right)^2 = r^2$$

$\therefore R^2 = r^2$

Since $0 \le R^2 \le 1$

$\therefore 0 \le r^2 \le 1$

or, $-1 \le r \le +1$.

Example 2.4. Find the value of R^2 from the following information and comment :

$\sum_{i=1}^{n} x_i y_i = 3347.60$, $\sum_{i=1}^{n} x_i^2 = 604.80$, $\sum_{i=1}^{n} y_i^2 = 19837$, $n = 20$, where $x_i = X_i - \overline{X}$ and $y_i = Y_i - \overline{Y}$

Solution : Since $R^2 = \dfrac{\hat{\beta}^2 \sum_{i-1}^{n} x_i^2}{\sum_{i-1}^{n} y_i^2}$ where $\hat{\beta} = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i-1}^{n} x_i^2}$

$$\therefore \hat{\beta}^2 = \left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}\right)^2 = \left(\frac{3347.60}{604.80}\right)^2 = (5.54)^2 = 30.69$$

$$\text{Now } R^2 = \frac{\hat{\beta}^2 \cdot \sum_{i-1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2} = \frac{30.69 \times 604.80}{19837} = \frac{18561.312}{19837} = 0.935$$

$\therefore R^2 = 0.935.$

This suggests that 93.5 percent of the changes in the sample observations of Y can be attributed to the variations of the fitted value of Y i.e., $\hat{Y}_i$ or we say that our regression line fits the given data well.

Thus R^2 measures the proportion of variations in the dependent variable that is explained by the independent variables.

Example 2.5. A sample of 20 observations corresponding to the regression model $Y_i = \alpha + \beta X_i + u_i$ where u_i is normally distributed with mean zero and unknown variance σ_u^2, gives the following data :

$$\sum_{i=1}^{n} Y_i = 21.9,\ \sum_{i=1}^{n}(Y_i-\bar{Y})^2 = 86.9,\ \sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}) = 106.4$$

$$\sum_{i=1}^{n} X_i = 186.2,\ \sum_{i=1}^{n}(X_i-\bar{X})^2 = 215.4,\ n = 20$$

Obtain the usual regression results.

Solution : On the basis of the given information we have to fit a linear relation between Y (dependent variable) and X (explanatory variable).

(i) **Estimation of $\hat{\alpha}$ and $\hat{\beta}$:**

We know that, $$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

$$\therefore \hat{\beta} = \frac{106.4}{215.4} = 0.494$$

and $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$ where $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{21.9}{20} = 1.095$ and $\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n} = \frac{186.2}{20} = 9.31$

$= 1.095 - 0.494 \times 9.31$

$= 1.095 - 4.60 = -3.505$

Thus we have, $\hat{\alpha} = -3.505$ and $\hat{\beta} = 0.494$ and our estimated regression line is :

$\hat{Y} = \hat{\alpha} + \hat{\beta}X \Rightarrow \hat{Y} = -3.505 + 0.494X$

(ii) **Estimation of variances :**

Since we know that, $$\text{var}(\hat{\alpha}) = \sigma_u^2 \left(\frac{\sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n} x_i^2} \right) \text{ and } \text{var}(\hat{\beta}) = \frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2}.$$

Here we see that σ_u^2 is not known and hence we replace it by its unbiased estimator $\hat{\sigma}_u^2 = \sum_{i=1}^{n} e_i^2 / (n-2)$.

Thus we have, $$\text{var}(\hat{\alpha}) = \hat{\sigma}_u^2 \left(\frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2} \right) \text{ and } \text{var}(\hat{\beta}) = \frac{\hat{\sigma}_u^2}{\sum_{i=1}^{n} x_i^2}.$$

Again we know that, $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} y_i^2 - \hat{\beta}^2 \sum_{i=1}^{n} x_i^2$

$\left[\text{Since } \sum_{i=1}^{n} y_i^2 = \hat{\beta}^2 \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} e_i^2 \text{ where } \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \text{ and } \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2\right]$

$\therefore \sum_{i=1}^{n} e_i^2 = 86.9 - (0.494)^2 \times 215.4 = 86.9 - 52.56 = 34.34$

Now $\hat{\sigma}_u^2 = \sum_{i=1}^{n} e_i^2 \Big/ (n-2) = \dfrac{34.34}{20-2} = \dfrac{34.34}{18} = 1.908$

Now $\text{var}(\hat{\alpha}) = \hat{\sigma}_u^2 \left(\dfrac{\sum_{i=1}^{n} X_i^2}{n \cdot \sum_{i=1}^{n} x_i^2} \right) = \dfrac{1.908 \times 1948.922}{20 \times 215.4} = 0.8631$

$\left[\because \sum_{i=1}^{n} (X_i - \bar{X})^2 = 215.4 \text{ or, } \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 = 215.4\right.$

or, $\sum_{i=1}^{n} X_i^2 = 215.4 + n\bar{X}^2 = 215.4 + 20 \times (9.3)^2$

$\left. = 215.4 + 1733.522 = 1948.922 \right]$

$\therefore \text{var}(\hat{\alpha}) = 0.8631$

Similarly, $\text{var}(\hat{\beta}) = \dfrac{\hat{\sigma}_u^2}{\sum_{i=1}^{n} x_i^2} = \dfrac{1.908}{215.4} = 0.0089$

Now, $SE(\hat{\alpha}) = \sqrt{\text{var}(\hat{\alpha})} = \sqrt{0.8631} = 0.929$

$SE(\hat{\beta}) = \sqrt{\text{var}(\hat{\beta})} = \sqrt{0.0089} = 0.094.$

(iii) Construction of confidence intervals :

Now we like to set up a confidence interval for α and β at (a) P = 0.95 (i.e., 5% level of significane) and (b) P = 0.99 (i.e., 1% level of significance)

In other words, we like to find the value of '*t*' that cuts off (a) 0.025 and (b) 0.005 of the area at the tail end of the distribution on both sides. From table value : $t_{0.025}$, $(n-2) = t_{0.025}$, 18 = 2.101 and $t_{0.005}$, $(n-2) = t_{0.005}$, 18 = 2.878

Therefore 95% confidence interval for α are : $\hat{\alpha} \pm t_{0.025}, (n-2)\ SE(\hat{\alpha})$ i.e., $P[\hat{\alpha} - t_{0.025},(n-2)\ SE(\hat{\alpha}) \le \alpha \le \hat{\alpha} + t_{0.025},(n-2)\ SE(\hat{\alpha})] = 0.95$ and 99% confidence interval for α would be : $\hat{\alpha} \pm t_{0.025},(n-2)\ SE(\hat{\alpha})$.

i.e., $P[\hat{\alpha} - t_{0.005}, (n-2)\ SE(\hat{\alpha}) \le \alpha \le \hat{\alpha} + t_{0.005}, (n-2)\ SE(\hat{\alpha})] = 0.99$

Therefore 95% confidence interval for α would be : $\hat{\alpha} \pm t_{0.025}, \ (n-2)\ SE(\hat{\alpha})$.

$\Rightarrow -3.505 \pm 2.101 \times 0.929$ or, -3.505 ± 1.9518.

Similarly, 99% confidence interval for α would be :

$-3.505 \pm 2.878 \times 0.929$ or, -3.505 ± 2.6736.

Similarly, 95% confidence interval of β are : $\hat{\beta} \pm t_{0.025}, \ (n-2)\ SE(\hat{\beta})$

i.e., $P[\hat{\beta} - t_{0.025}, (n-2)\ SE(\hat{\beta}) \le \beta \le \hat{\beta} + t_{0.025}, (n-2)\ SE(\hat{\beta})] = 0.95$.

and 99% confidence interval for β are : $\hat{\beta} \pm t_{0.005}, (n-2)\ SE(\hat{\beta})$

i.e., $P[\hat{\beta} - t_{0.005}, (n-2) SE(\hat{\beta}) \le \beta \le \hat{\beta} + t_{0.005}, (n-2) SE(\hat{\beta})] = 0.99$

Thus 95% confidence internal for β would be : $\hat{\beta} \pm t_{0.025}, \ (n-2) SE(\hat{\beta})$

or, $0.494 \pm 2.101 \times 0.094$

or, 0.494 ± 0.1974

where $\hat{\beta} = 0.494$

$t_{0.025}, \ (n-2) = t_{0.025}, \ 18 = 2.101$

$SE(\hat{\beta}) = 0.094$.

(iv) **Hypothesis testing :** Suppose we like to test the null hypothesis $H_0 : \beta = 0$ against the alternative hypothesis $H_1 : \beta \neq 0$. Now on the basis of the given sample $H_0 : \beta = 0$ will be rejected at 5% level of significance if

$$|t_{n-2}| = \left| \frac{\hat{\beta}}{SE(\hat{\beta})} \text{(observed)} \right| > t_{0.025}, \ (n-2) \text{ (table value)}$$

and will be accepted otherwise.

$$\text{Here } t_{n-2} = \frac{\hat{\beta}}{SE(\hat{\beta})} = \frac{0.494}{0.094} = 5.255 \text{ (where } n = 20).$$

Thus we see that, $|t_{n-2}| = 5.255 > t_{0.025}, \ 18$ (=2.101) and hence $H_0 : \beta = 0$ is rejected (alternative $H_1 : \beta \neq 0$ is accepted) at 5% level of significance. So, the hypothesis of no relationship between X and Y is to be rejected at 5% level of significance. Similarly, it can be tested for 1% level of significance.

2.15. Results of Regression Analysis

The results of regression analysis are generally presented in a conventional format. It is not sufficient merely to report the estimates of α and β. In practice we report regression coefficients together with their standard errors and the value of R^2. It has become customary to present the estimated equation with standard errors placed in parentheses below the estimated parameter values. These results are supplemented by R^2, the value of which is written on right hand side of the estimated regression equation.

In terms of our earlier example (Example 2.5) the estimated regression results can be written as :

$Y = -3.505 + 0.494\ X \qquad R^2 = 0.6048.$

SE : (0.929) (0.094)

Here $\hat{\alpha} = -3.505$, $\hat{\beta} = 0.494$, $SE(\hat{\alpha}) = 0.929$, $SE(\hat{\beta}) = 0.094$,

and $$R^2 = \frac{\hat{\beta}^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2} = \frac{(0.494)^2 \times 215.4}{86.9} = \frac{52.5653}{86.9} = 0.6048$$

This suggests that variations in 60.48 percent of the sample observations of Y can be attributed to the variations of the fitted value of Y or we say that our regression line fits the given data moderately (not very well).

Some econometricians report the t-ratios of the estimated coefficients in place of standard errors. This way of presentation makes the testing of hypothesis easier and direct.

Thus the other form of presentation of results for the earlier example would be :

$Y = \quad -3.505 + 0.494\,X \qquad R^2 = 0.6048$

t ratios : $(-3.772)\ (5.255)$

where $\hat{\alpha} = -3.505$, $\hat{\beta} = 0.494$

$$\frac{\hat{\alpha}}{SE(\hat{\alpha})} = -3.772, \quad \frac{\hat{\beta}}{SE(\hat{\beta})} = 5.255.$$

Example 2.6. Suppose that Mr. X estimates a consumption function and obtains the results :

$\hat{C} = \quad 15 + 0.81Y_d, \qquad n = 19$

t ratios : $(3.1)\ (18.7) \qquad R^2 = 0.99$

$\hat{C}$ is consumption ; Y_d is disposable income, the numbers in parentheses are t-ratios.

(a) Test the significance of Y_d statistically using t-ratios.

(b) Determine the estimated standard derivations of the parameter estimators.

(c) Construct a 95 percent confidence interval for the coefficient of Y_d.

Solution : It is a formal consumption function of the Keynesian type i.e., $C = a + bY_d$ where a = autonomous part of consumption and b = Marginal propensity to consume. By assumption in the existing theory ($a > 0$, $0 < b < 1$). The estimated relation/regression results are given here as :

$\hat{C} = \quad 15 + 0.81Y_d \qquad n = 19$

t ratios : $(3.1)\ (18.7) \qquad R^2 = 0.99$

This shows that $\hat{a} = 15$, $\hat{b} = 0.81$,

$$\frac{\hat{a}}{SE(\hat{a})} = 3.1, \text{ (t-ratio) and } \frac{\hat{b}}{SE(\hat{b})} = 18.7 \text{ (t -ratio)}$$

n = no. of data points (sample size = 19)

R^2 = Square of multiple correlation coefficient

$$= \frac{\text{ESS}}{\text{TSS}} = \frac{\text{Explained variation in } C}{\text{Total variation}} = \frac{\text{var}(\hat{C})}{\text{var}(C)}.$$

$$= \frac{\hat{b}^2 \sum_{i=1}^{n} y_{d_i}^2}{\sum_{i=1}^{n} c_i^2} \text{ where } y_{d_i} = Y_{d_i} - \bar{Y}_d \quad c_i = C_i - \bar{C}$$

$$= 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} c_i^2} \text{ where } \sum_{i=1}^{n} c_i^2 = \sum_{i=1}^{n} e_i^2 + \hat{b}^2 \sum_{i=1}^{n} y_{d_i}^2$$

Here $R^2 = 0.99 = \frac{\text{ESS}}{\text{TSS}} = \frac{99}{100}$.

This means that 99 percent of the variations in sample observations of C can be attributed to the variations of the fitted value of C i.e., $\hat{C}$. Thus we say that our regression line fits the given data well. Out of 100% variation in consumption our regression relation can explain 99% variation in consumption.

(a) We have to test the null hypothesis $H_0 : b = 0$ (no relation between C and Y_d) against the alternative hypothesis $H_1 : b \neq 0$.

The appropriate test statistic under $H_0 : b = 0$ would be :

$\frac{\hat{b}}{SE(\hat{b})}$ which follows a t distribution with $(n - 2)$ degrees of freedom.

i.e., $t_{n-2} = t = \frac{\hat{b}}{SE(\hat{b})} = 18.7$ (given).

Now at 5% level of significance $H_0 : b = 0$ (no relation between C and Y_d) will be accepted if $-t_{0.025,n-2} \leq t \leq t_{0.025,n-2}$

and will be rejected otherwise.

From table value we get,

$t_{0.025,n-2} = t_{0.025,17} = 2.110$

$(n = 19$, given)

Thus we see that observed $t = \frac{\hat{b}}{SE(\hat{b})} = 18.7$ does not lie in the interval -2.110 and 2.110 and hence the null hypothesis is rejected and the alternative is accepted. This means that there exists a relation between consumption (C) and disposable income (Y_d). Hence the relation is statistically significant.

(b) We have to find $SE(\hat{a})$ and $SE(\hat{b})$.

Since for $\hat{a}, t = \frac{\hat{a}}{SE(\hat{a})} = 3.1$ (given)

and $\hat{a} = 15$, $\therefore 3.1 = \frac{15}{SE(\hat{a})}$ or, $SE(\hat{a}) = \frac{15}{3.1} = 4.8387$

Similarly, for $\hat{b}$, $t = \frac{\hat{b}}{SE(\hat{b})} = 18.7$ (given)

(and $\hat{b} = 0.81$)

or, $18.7 = \frac{0.81}{SE(\hat{b})}$ or, $SE(\hat{b}) = \frac{0.81}{18.7} = 0.0433$

Thus the estimated standard deviations of the parameter estimators are : $SE(\hat{a}) = 4.8387$, $SE(\hat{b}) = 0.0433$.

(c) We have to construct 95 per cent confidence interval for the coefficient of Y_d (i.e., b).

This will be given by :

$\hat{b} \pm t_{0.025, n-2} SE(\hat{b})$

or, $0.81 \pm 2.110 \times 0.0433$

or, 0.81 ± 0.0913

i.e., between 0.7187 and 0.9013

This means that the coefficient of Y_d will lie between 0.7187 and 0.9013.

since

$t_{0.025, n-2}$
$= t_{0.025, 17} = 2.110$

$SE(\hat{b}) = 0.0433$

$\hat{b} = 0.81$

Example 2.7. The following table shows data on Labour-hours of work and output for 10 workers :

X (labour hours of work) :	10	7	10	5	8	8	6	7	9	10
Y (output) :	11	10	12	6	10	7	9	10	11	10

(i) Assuming a linear regression of the form $Y = \alpha + \beta X + u \; (u \sim N(0, \sigma_u^2))$, find the OLS estimators of α and β (i.e., $\hat{\alpha}$ and $\hat{\beta}$) and the estimated regression line $\hat{Y} = \hat{\alpha} + \hat{\beta} X$.

(ii) Find var($\hat{\alpha}$), var($\hat{\beta}$), $SE(\hat{\alpha})$ and $SE(\hat{\beta})$.

(iii) Find the value of $\sum_{i=1}^{n} e_i^2$ and $\hat{\sigma}_u^2 = \sum_{i=1}^{n} e_i^2 \Big/ (n-2)$

(iv) Find the value of R^2

(v) Construct 95% confidence intervals of α, β and σ_u^2.

(vi) Test the null hypothesis $H_0 : \beta = 1.35$ against
(a) $H_1 : \beta \neq 1.35$, (b) $H_1 : \beta > 1.35$, (c) $H_1 : \beta < 1.35$

Solution : **Calculations for the Regression**

Observations	X_i	Y_i	$x_i = X_i - \bar{X}$	$y_i = Y_i - \bar{Y}$	x_i^2	$x_i y_i$	$\hat{Y}_i$	$e_i = Y_i - \hat{Y}_i$
1	10	11	2	1.4	4	2.8	11.10	− 0.1
2	7	10	− 1	0.4	1	− 0.4	8.85	1.15
3	10	12	2	2.4	4	4.8	11.10	0.9
4	5	6	− 3	− 3.6	9	10.8	7.35	− 1.35
5	8	10	0	0.4	0	0	9.60	0.4
6	8	7	0	− 2.6	0	0	9.60	− 2.6
7	6	9	− 2	− 0.6	4	1.2	8.10	0.9
8	7	10	− 1	0.4	1	− 0.4	8.85	1.15
9	9	11	1	1.4	1	1.4	10.35	0.65
10	10	10	2	0.4	4	0.8	11.10	− 1.1
Total —	$\sum_{i=1}^{n} X_i = 80$	$\sum_{i=1}^{n} Y_i = 96$	$\sum_{i=1}^{n} x_i = 0$	$\sum_{i=1}^{n} y_i = 0$	$\sum_{i=1}^{n} x_i^2 = 28$	$\sum_{i=1}^{n} x_i y_i = 21$	$\sum_{i=1}^{n} \hat{Y}_i = 96$	$\sum_{i=1}^{n} e_i = 0$

Here $n = 10$, $\bar{X} = \left(\sum_{i=1}^{n} X_i \Big/ n\right) = \frac{80}{10} = 8$, $\bar{Y} = \left(\sum_{i=1}^{n} Y_i \Big/ n\right) = \frac{96}{10} = 9.6$,

$$\sum_{i=1}^{n} e_i^2 = (0.1)^2 + (1.15)^2 + \ldots + (-1.1)^2 = 14.65$$

and $\sum_{i=1}^{n} X_i^2 = 10^2 + 7^2 + \ldots + 10^2 = 668$

(i) The OLS estimators of α and β are given by $\hat{\alpha}$ and $\hat{\beta}$ where

$$\hat{\beta} = \sum_{i=1}^{n} x_i y_i \Big/ \sum_{i=1}^{n} x_i^2 = \frac{21}{28} = 0.75$$

and $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 9.6 - 0.75 \times 8 = 9.6 - 6 = 3.6$

$\therefore$ $\hat{\alpha} = 3.6$ and $\hat{\beta} = 0.75$.

The estimated regression line/equation becomes

$\hat{Y} = \hat{\alpha} + \hat{\beta}X$ or, $\hat{Y} = 3.6 + 0.75X$. This equation is now used to find the values of $\hat{Y}_i$ corresponding to different values of X_i. The values of $\hat{Y}_i$ are given in the above table showing calculations for regression.

Here the regression coefficient $\hat{\beta} = 0.75$ measures the marginal productivity of labour. The intercept $\hat{\alpha} = 3.6$ means that output will be 3.6 units when labour hours of work is zero.

(ii) We have to find $\operatorname{var}(\hat{\alpha})$, $\operatorname{var}(\hat{\beta})$, $SE(\hat{\alpha})$ and $SE(\hat{\beta})$.

Since we know that $\operatorname{var}(\hat{\alpha}) = \sigma_u^2 \left(\dfrac{\sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} x_i^2} \right)$ and $\operatorname{var}(\hat{\beta}) = \sigma_u^2 \Big/ \sum_{i=1}^{n} x_i^2$.

Since σ_u^2 is unknown it is replaced by its unbiased estimator,

$$\hat{\sigma}_u^2 = \sum_{i=1}^{n} e_i^2 \Big/ (n-2) = \frac{14.65}{10-2} = \frac{14.65}{8} = 1.8312$$

$$\text{Now } \operatorname{var}(\hat{\alpha}) = \frac{\hat{\sigma}_u^2 \sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} x_i^2} = \frac{1.8312 \times 668}{10 \times 28} = \frac{1223.2416}{280} = 4.3687$$

and $SE(\hat{\alpha}) = \sqrt{\operatorname{var}(\hat{\alpha})} = \sqrt{4.3687} = 2.090$

Again, $\operatorname{var}(\hat{\beta}) = \dfrac{\hat{\sigma}_u^2}{\sum_{i=1}^{n} x_i^2} = \dfrac{1.8312}{28} = 0.0654$ and $SE(\hat{\beta}) = \sqrt{\operatorname{var}(\hat{\beta})} = \sqrt{0.0654} = 0.256$

(iii) We have to find out the values of $\sum_{i=1}^{n} e_i^2$ and $\hat{\sigma}_u^2 = \sum_{i=1}^{n} e_i^2 \Big/ (n-2)$.

Here $\sum_{i=1}^{n} e_i^2 = (-0.1)^2 + (1.15)^2 + (0.9)^2 + \ldots + (0.65)^2 + (-1.1)^2 = 14.65$ [e_i values are taken from last column of calculation table]

and $\hat{\sigma}_u^2 = \sum_{i=1}^{n} e_i^2 \Big/ (n-2) = \dfrac{14.65}{10-2} = \dfrac{14.65}{8} = 1.8312$

(iv) We have to find out the value of the coefficient of determination, R^2.

Since we know that $R^2 = \dfrac{ESS}{TSS} = \dfrac{\hat{\beta}^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2}$ [Since $\sum_{i=1}^{n} y_i^2 = \hat{\beta}^2 \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} e_i^2$ i.e., $TSS = ESS + RSS$]

$$\therefore \sum_{i=1}^{n} y_i^2 = \hat{\beta}^2 \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} e_i^2$$

$$= (0.75)^2 \times 28 + 14.65 = 15.75 + 14.65 = 30.40$$

$$\therefore R^2 = \frac{\hat{\beta}^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2} = \frac{(0.75)^2 \times 28}{30.40} = \frac{15.75}{30.40} = 0.51$$

This suggests that 51 percent of the variations in the sample observations of Y can be attributed to the variations of the fitted value of Y i.e., $\hat{Y}_i$. Here we see that our regression line fits the given data moderately (not very well).

From the above results we can write our regression results as follows :

$\hat{y} = 3.6 + 0.75\,X,\ R^2 = 0.51$

(2.090) (0.256) [SE values in brackets]

Alternatively, $\hat{y} = 3.6 + 0.75\,X,\ R^2 = 0.51$

(1.7224) (2.930) [t values in brackets]

(v) **95% confidence intervals of α, β and σ_u^2**

(a) 95% confidence interval for α is given by,

$P[\hat{\alpha} - t_{0.025,n-2} \cdot SE(\hat{\alpha}) \le \alpha \le \hat{\alpha} + t_{0.025,n-2} \cdot SE(\hat{\alpha})] = 0.95$

i.e., 95% confidence limits of α are :

$\hat{\alpha} \pm t_{0.025,n-2} \cdot SE(\hat{\alpha})$

or, $3.6 \pm 2.306 \times 2.090$

or, 3.6 ± 4.820

$\Rightarrow -1.22$ and 8.42

Since

$t_{0.025,n-2} = t_{0.025,8}$
$= 2.306$ (from table value)

So, 95% confidence limits for α are -1.22 and 8.42.

(b) 95% confidence interval for β is given by,

$$P[\hat{\beta} - t_{0.025,n-2} \cdot SE(\hat{\beta}) \leq \beta \leq \hat{\beta} + t_{0.025,n-2} \cdot SE(\hat{\beta})] = 0.95$$

i.e., 95% confidence limits of β are :

$\hat{\beta} \pm t_{0.025,n-2} \cdot SE(\hat{\beta})$

or, $0.75 \pm t_{0.025,8} \cdot SE(\hat{\beta})$

or, $0.75 \pm 2.306 \times 0.256$

or, 0.75 ± 0.590

$\Rightarrow 0.16$ and 1.34

Since $t_{0.025,n-2} = t_{0.025,8} = 2.306$
(From table value)

$\therefore$ 95% confidence limits of β are 0.16 and 1.34.

(c) Since we know that 100 $(1 - \alpha)$% confidence interval for σ_u^2 is given by,

$$P\left[(n-2) \cdot \frac{\hat{\sigma}_u^2}{\chi^2_{\alpha/2,n-2}} \leq \sigma_u^2 \leq (n-2)\frac{\hat{\sigma}_u^2}{\chi^2_{1-\alpha/2,n-2}}\right] = 1 - \alpha$$

Here $n = 10$, $\alpha = 0.05$, $\hat{\sigma}_u^2 = \sum_{i=1}^{n} e_i^2 \Big/ (n-2) = 1.8312$, $X^2_{\alpha/2,n-2} = \chi^2_{0.025,8} = 17.535$ (Table value)

and $\chi^2_{1-\alpha/2,n-2} = \chi^2_{0.975,8} = 2.180$ (Table value)

so, 95% confidence interval for σ_u^2 would be

$$P\left[8 \times \frac{1.8312}{17.535} \leq \sigma_u^2 \leq 8 \times \frac{1.8312}{2.180}\right] = 0.95$$

or, $P[0.84 \leq \sigma_u^2 \leq 6.72] = 0.95$

$\therefore$ 95% confidence limits of σ_u^2 are 0.84 and 6.72.

(vi) To test the null hypothesis $H_0 : \beta = 1.35$ against the alternative hypothesis $H_1 : \beta \neq 1.35$ the appropriate test statistic would be, $t_{n-2} = \dfrac{\hat{\beta} - \beta}{SE(\hat{\beta})}$.

Now on the basis of the sample data $H_0 : \beta = 1.35$ will be rejected at 5% level of significance if

$|t_{n-2}| = \dfrac{\hat{\beta} - \beta}{SE(\hat{\beta})}$ (observed) > $t_{0.025,\ n-2}$ (Table value)

and will be accepted otherwise.

Here $t_{n-2} = \dfrac{\hat{\beta}-\beta}{SE(\hat{\beta})} = \dfrac{0.75-1.35}{0.256} = \dfrac{-0.6}{0.256} = -2.343$

$\therefore t_{n-2} = -2.343$ and $|t_{n-2}| = 2.343$

From the table value we get $t_{0.025,\, n-2} = t_{0.025,8} = 2.306$.

Thus, $|t_{n-2}| = 2.343 > t_{0.025,8} = 2.306$. This means that on the basis of sample data, $H_0 : \beta = 1.35$ is rejected at 5% level of significance.

(b) The null hypothesis $H_0 : \beta = 1.35$ will be rejected against the alternative $H_1 : \beta > 1.35$ at 100α% level of significance if for the given sample

$$t_{n-2} \text{ (observed)} = \frac{\hat{\beta}-\beta}{SE(\hat{\beta})} > t_{\alpha,\, n-2} \text{ (table value)}$$

and will be accepted otherwise. Here $\alpha = 0.05$, $n = 10$

$$\text{and } t_{n-2} \text{ (observed)} = \frac{\hat{\beta}-\beta}{SE(\hat{\beta})} = -2.343 < t_{0.5,8} = 1.860.$$

So, $H_0 : \beta = 1.35$ is accepted (hence in significant) at 5% level of significance.

(c) The null hpothesis, $H_0 = 1.35$ will be rejected against the alternative $H_1 : \beta < 1.35$ for the given sample at 100α% level of significance if

$$t_{n-2} \text{ (observed)} = \frac{\hat{\beta}-\beta}{SE(\hat{\beta})} < t_{\alpha,\, n-2} \text{ (table value)}$$

and will be accepted otherwise.

Here, t_{n-2} (observed) $= \dfrac{\hat{\beta}-\beta}{SE(\hat{\beta})} = \dfrac{0.75-1.35}{0.256} = -2.343$

$< t_{0.05,8} = -1.860$ Here $\alpha = 0.05$, $n = 10$

This clearly shows that the null hypothesis $H_0 : \beta = 1.35$ is rejected(significant) at 5% level of significance.

2.16. Analysis of Variance for the Simple Linear Regression Model

Yet another item that is often presented in connection with the simple linear regression model is the **analysis of variance.** This is the breakdown of the total sum of squares (TSS) into explained sum of squares (ESS) and the residual sum of squares (RSS). The purpose of presenting the table is to test the significance of the explained sum of squares. In this case this amounts to testing the significance of β.

In regression analysis, we minimise the square deviations from mean and it has been proved (see Section 2.14) that,

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y}_i)^2 + \sum_{i=1}^{n}(Y_i - \hat{Y})^2 \text{ or, } \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} e_i^2$$

That is, Total variation = Explained variation + Unexplained variation or Residual variance or, TSS = ESS + RSS with degrees of freedom : $n - 1 = (K - 1) + (n - K)$ where

n = total number of observations (given) and

K = number of parameters to be estimated.

If the model is a two variable regression model then

$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$ and $Y_i = \alpha + \beta X_i + e_i$, e_i is being the error of estimate.

Now $e_i = Y_i - \hat{Y}_i$

$$\text{or, } Y_i = \hat{Y}_i + e_i \text{ or, } Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + e_i$$

$$\text{or, } \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}\left[(\hat{Y}_i - \bar{Y}) + e_i\right]^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + 2\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})e_i + \sum_{i=1}^{n}e_i^2$$

It can be proved that (see Section 2.14)

$$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})e_i = 0 \text{ and } \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = \hat{\beta}^2\sum_{i=1}^{n}(X_i - \bar{X})^2 = \hat{\beta}^2\sum_{i=1}^{n}x_i^2$$

and hence we have,

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \hat{\beta}^2\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}e_i^2 \text{ or, } \sum_{i=1}^{n}y_i^2 = \hat{\beta}^2\sum_{i=1}^{n}x_i^2 + \sum_{i=1}^{n}e_i^2$$

i.e., TSS = ESS + RSS

with d.f. $n - 1 = (K - 1) + (n - K)$. Here $K = 2$ as there are two parameters α & β. Thus we see that total variations are split into explained (by explanatory variables) and unexplained (error terms) variations against *between and within variations* in the case of analysis of variace procedure. This suggests that we can compile an analysis of variance type of table for the regression analysis also in order to judge the overall significance of the regression results.

ANOVA TABLE

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	F	
				Observed	Tabulated
Explained (between)	ESS = $\hat{\beta}^2\sum_{i=1}^{n}x_i^2$	$K - 1$	ESS/(K − 1) = MSE	$F = \frac{\text{MSE}}{\text{MSR}}$ with df $= (K - 1)$, $(n - K)$	
Residual (within)	RSS = $\sum_{i=1}^{n}e_i^2$	$n - K$	RSS/$(n - K)$ = MSR		
Total	TSS = $\sum_{i=1}^{n}y_i^2$	$n - 1$			

Here $K = 2$ as the model is a two variable regression model and two parameters are involved.

To test the null hypothesis concerning the significance of the regression (i.e., whether the explanatory variable is a significant factor of the variation in Y), we perform F test by computing F-ratio through dividing mean sum of squares of explained variations to that of residual. In testing $H_0 : \beta = 0$ against the alternative $H_1 : \beta \neq 0$ we may use the test statistic,

$$F^* = \frac{\text{MSE}}{\text{MSR}} = \frac{\hat{\beta}^2 \sum x_i^2 / (K-1)}{\sum e_i^2 / (n-K)}$$

$$= \frac{\hat{\beta}^2 \sum x_i^2 / 1}{\sum e_i^2 / (n-2)} \text{ with d.f.} = 1, (n-2), \text{ (Here } K = 2).$$

Now we have to compare $F^*_{1,n-2}$ with the table value of F with d.f. = 1, $n-2$. If it is found that $F^* > F_{\alpha,1,n-2}$ (table), we reject the null hypothesis at $100\alpha\%$ level of significance ($\alpha = 0.01$ or 0.05) and accept otherwise. The relation will be significant if H_0 is rejected.

It should be noted that,

$$F^* = \frac{\text{MSE}}{\text{MSR}} = \frac{\hat{\beta}^2 \sum x_i^2}{\sum e_i^2 / (n-2)} = \frac{(n-2)\hat{\beta}^2 \sum x_i^2}{(1-R^2)\sum y_i^2}$$

$$\left[\text{Now } \frac{\hat{\beta}^2 \sum x_i^2}{\sum y_i^2} = \frac{\sum y_i^2 - \sum e_i^2}{\sum y_i^2} = 1 - \frac{\sum e_i^2}{\sum y_i^2}\right.$$

$$\left.\text{But } \frac{\sum e_i^2}{\sum y_i^2} = 1 - R^2 \quad \text{or, } \sum e_i^2 = (1-R^2)\sum y_i^2\right]$$

$$\text{Therefore, } F^* = \frac{(n-2)\left\{\dfrac{\hat{\beta}^2 \sum x_i^2}{\sum y_i^2}\right\}}{1-R^2} = \frac{(n-2)R^2}{1-R^2}$$

which, on generalisation becomes $\dfrac{R^2/(K-1)}{(1-R^2)/(n-K)}$ where we have K parameters.

Furthermore, we know that,

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})} = \frac{\hat{\beta}}{\sqrt{\text{var}(\hat{\beta})}}. \text{ But } \text{var}(\hat{\beta}) = \frac{\hat{\sigma}_u^2}{\sum x_i^2} = \frac{\sum e_i^2 / (n-2)}{\sum x_i^2}$$

$$\therefore t^2 = \frac{\hat{\beta}^2}{\text{var}(\hat{\beta})} = \frac{\hat{\beta}^2}{\left\{\sum e_i^2/(n-2)\right\}\left(\dfrac{1}{\sum x_i^2}\right)} \quad \text{or, } t^2 = \frac{\hat{\beta}^2 \sum x_i^2}{\sum e_i^2 / (n-2)} = F^*.$$

This means that t and F tests are formally equivalent, the relation between the two being : $t^2 = F$.

Example 2.8. Let us consider the following data to construct the analysis of variance table for a simple regression model : $Y_i = \alpha + \beta X_i + u_i$

$$\text{Given : } \sum_{i=1}^{n} Y_i = 21.9,\ \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = 86.9,\ \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = 106.4,$$

$$\sum_{i=1}^{n} X_i = 186.2,\ \sum_{i=1}^{n} (X_i - \bar{X})^2 = 215.4,\ n = 20$$

Solution : [See Example 2.5]

The OLS estimators of α and β can be obtained as follows :

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{106.4}{215.4} = 0.494$$

and $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$

$= 1.095 - 0.494 \times 9.31 = 1.095 - 4.60 = -3.505$

where $\bar{Y} = \left(\sum_{i=1}^{n} Y_i \Big/ n\right) = \frac{21.9}{20} = 1.095$ and $\bar{X} = \left(\sum_{i=1}^{n} X_i \Big/ n\right) = \frac{186.2}{20} = 9.31$

The estimated regression results are :

$\hat{y} = - 3.505 + 0.494\, X,\ R^2 = 0.6048$

t ratios : (-3.772) (5.255) where $\hat{\alpha} = -3.505$, $\hat{\beta} = 0.494$

$$SE(\hat{\alpha}) = 0.929,\ SE(\hat{\beta}) = 0.094,\ \frac{\hat{\alpha}}{SE(\hat{\alpha})} = -3.772,\ \frac{\hat{\beta}}{SE(\hat{\beta})} = 5.255$$

$$R^2 = \frac{\hat{\beta}^2 \sum x_i^2}{\sum y_i^2} = \frac{(0.494)^2 \times 215.4}{86.9} = 0.6048$$

Now $\sum e_i^2 = \sum y_i^2 - \hat{\beta}^2 \sum_{i=1}^{n} x_i^2 = 86.9 - (0.494)^2 \times 215.4 = 34.34.$

Now for testing $H_0 : \beta = 0$

against the alternative $H_1 : \beta \neq 0$, we may use the ANOVA Table.

ANOVA TABLE

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	F	
				Observed	Table value at 1% and 5%
Explained (between)	$ESS = \hat{\beta}^2 \sum_{i=1}^{n} x_i^2$ $= 52.56$	$K - 1$ $= 2 - 1 = 1$	$MSE = \frac{ESS}{K-1}$ $= \frac{52.56}{1}$	$F^* = \frac{MSE}{MSR}$ $= \frac{52.56}{1.908}$ $= 27.55$	$F_{0.01,1,18}$ $= 8.29$
Residual (within)	$RSS = \sum_{i=1}^{n} e_i^2$ $= 34.34$	$n - K$ $= 20 - 2 = 18$	$MSR = \frac{RSS}{n-K}$ $= \frac{34.34}{18}$ $= 1.908$	with df 1, 18	$F_{0.05,1,18}$ $= 4.41$
Total	$TSS = \sum_{i=1}^{n} y_i^2$ $= 86.9$	$n - 1$ $= 20 - 1 = 19$			

Here we see that the observed $F^* = 27.55$ is much larger than table $F_{0.01}$; 1, 18 = 8.29 and $F_{0.05}$; 1, 18 = 4.41. This means that $H_0 : \beta = 0$ is rejected both at 5% and 1% levels of significance. Hence, we reject the null hypothesis and accept that the regression is significant, that is, X is a significant explanatory factor of the variation in Y.

2.17. Testing the Equality between Coefficients Obtained from Different Regressions or Different Samples

Sometimes we may have to estimate a regression equation separately for several sets of data and we may have to test whether some or all the parameters are the same for all different sets of data.

Suppose, we have two samples on the variables Y and X containing n_1 observations for first set and, Y and X containing n_2 observations for second set. We may obtain two estimates of the same relationship for these two samples ;

$$\hat{Y}_1 = \hat{\alpha}_1 + \hat{\beta}_1 X$$

and $\hat{Y}_2 = \hat{\alpha}_2 + \hat{\beta}_2 X$

Now our problem is to examine whether these two estimated relations differ significantly. If yes, then we may conclude that the relationship changes from one sample to the other.

For example, suppose that we have the data on consumption and disposable income for the two periods 1990-1999 and 2000-2019. We estimate the consumption functions separately for these periods. Then we may be interested to examine whether the functions are statistically significant or whether the MPCs significantly differ or not.

Prof. G. C. Chow suggested F-test to answer these questions. For the purpose of the test we have to consider the following steps :

Step 1 : We have to fit the pooled regression with number of observations $= (n_1 + n_2)$ and estimate unexplained variation $\sum e_p^2$ i.e., to obtain $\hat{Y}_p = \hat{a}_0 + \hat{a}_1 X$

and $\sum e_p^2 = \sum y_p^2 - \sum \hat{y}_p^2$ with $(n_1 + n_2 - 2K)$ degrees of freedom, p stands for pooled data.

Step 2 : Next we fit the regression for each sample separately.

For first sample : $\hat{Y}_1 = \hat{\alpha}_1 + \hat{\beta}_1 X$ and $\sum e_1^2 = \sum y_1^2 - \sum \hat{y}_1^2$

For second sample : $\hat{Y}_2 = \hat{\alpha}_2 + \hat{\beta}_2 X$ and $\sum e_2^2 = \sum y_2^2 - \sum \hat{y}_2^2$

Step 3 : Next we have to compute F^* ratio as follows :

$$F^* = \frac{\left\{\sum e_p^2 - \left(\sum e_1^2 + \sum e_2^2\right)\right\} / K}{\left(\sum e_1^2 + \sum e_2^2\right) / (n_1 + n_2 - 2K)} \quad \text{with d.f. } K, (n_1 + n_2 - 2K)$$

Next we have to test the null hypothesis :

$H_0 : \hat{a}_1 = \hat{\beta}_1 = \hat{\beta}_2$ against the alternative H_1 : H_0 is not correct.

If $F^* > F_{0.05}$, we reject the null hypothesis at 5% level of significance.

In particular, if the pooled results are not given then F^* ratio can be obtained as follows :

$$F^* = \frac{\sum e_1^2 / (n_1 - 2)}{\sum e_2^2 / (n_2 - 2)} \text{ with d.f : } \{(n_1 - 2), (n_2 - 2)\}$$

Example 2.9. In order to test the null hypothesis that there is no difference in the MPC (Marginal propensity to consume) of 'manual workers' and 'white-colour employees' a research team estimated the following consumption functions :

Manual workers : Sample size $n_1 = 35$

$$\hat{C}_1 = 120 + 0.90Y \quad r_1^2 = 0.92, \sum (C_1 - \bar{C}_1)^2 = 3,251$$

(32) (5.6)

(The numbers in brackets are the t values for the regression coefficients).

White-colour employees : Sample size $n_2 = 30$

$$\hat{C}_2 = 160 + 0.82Y \quad r_2^2 = 0.95, \sum (C_2 - \bar{C}_2)^2 = 4,532$$

(23) (8.5)

(The numbers in brackets are the 't' values for the regression coefficients).

Combined sample consumption function :

sample size $n = n_1 + n_2 = 30 + 35 = 65$

$$\hat{C} = 250 + 0.70Y \quad r^2 = 0.92, \sum e_p^2 = 16,320$$

(5.3) (6.2)

On the basis of the above results, can we accept the hypothesis that there is no differences between the MPCs of the two groups ? (use a 5 per cent level of significance).

Solution : Suppose the estimated consumption function for the first sample is $\hat{C}_1 = \hat{\alpha}_1 + \hat{\beta}_1 Y$ and for the second sample $\hat{C}_2 = \hat{\alpha}_2 + \hat{\beta}_2 Y$

and the pooled estimated consumption function is $\hat{C}_p = \hat{a}_0 + \hat{a}_1 Y$. We have to test the null hypothesis

$H_0 : \hat{a}_1 = \hat{\beta}_1 = \hat{\beta}_2$ against the alternative $H_1 : H_0$ is not true

(where $\hat{\beta}_1$ = MPC of the first sample

$\hat{\beta}_2$ = MPC of the second sample

$\hat{a}_1$ = MPC of the pooled sample)

The appropriate test statistic will be

$$F^* = \frac{\left\{\sum e_p^2 - \left(\sum e_1^2 + \sum e_2^2\right)\right\}/K}{\left(\sum e_1^2 + \sum e_2^2\right)/(n_1 + n_2 - 2K)} \text{ with}$$

$d.f = \{K, (n_1 + n_2 - 2K)\}$

Now H_0 : will be rejected (significant) if $F^* > F_{0.05}$; $K, (n_1 + n_2 - 2K)$ and will be accepted otherwise.

Now on the basis of our given information we see that :

From first sample :

$$n_1 = 35,\ r_1^2 = R_1^2 = 0.92,\ \sum (C_1 - \bar{C}_1)^2 = \sum c_1^2 = 3{,}251$$

$$K = 2,\ \hat{\beta}_1 = 0.90 \ \therefore \sum e_1^2 = (1 - R_1^2)\sum c_1^2 = (1 - 0.92) \times 3{,}251 = 260.08$$

$$\left[\text{since } R^2 = 1 - \frac{\sum e_i^2}{\sum y_i^2} \therefore \frac{\sum e_i^2}{\sum y_i^2} = 1 - R^2 \text{ or, } \sum e_i^2 = (1 - R^2)\sum y_i^2\right]$$

From second sample : $n_2 = 30$, $\hat{\beta}_2 = 0.82$, $r_2^2 = R_2^2 = 0.95$

$$\sum (C_2 - \bar{C}_2)^2 = \sum c_2^2 = 4{,}532, K = 2$$

and $\sum e_2^2 = (1 - R_2^2)\sum c_2^2 = (1 - 0.95) \times 4{,}532 = 226.6$

From the pooled sample : $n = n_1 + n_2 = 35 + 30 = 65$,

$\hat{a}_1 = 0.70,\ r^2 = R^2 = 0.92,\ \sum e_p^2 = 16{,}320$

Thus we have, $\dfrac{\left\{\sum e_p^2 - \left(\sum e_1^2 + \sum e_2^2\right)\right\}/K}{\left(\sum e_1^2 + \sum e_2^2\right)/(n_1 + n_2 - 2K)}$

$$= \frac{\{16{,}320 - (260.08 + 226.6)\}/2}{(260.08 + 226.6)/(35 + 30 - 2 \times 2)} \text{ with d.f. : } (K,\ n_1 + n_2 - 2K)$$

$$= \frac{(16{,}320 - 486.86)/2}{486.68/61} = \frac{7916.66}{7.9783} = 992.27$$

$\therefore F^* = 992.27$ with d.f : (2, 61).

From the table value we see that $F_{0.05,2,61}=3.148$

Thus we see that F^* is much larger than the table value (3.148) at 5% level of significance. Hence the null hypothesis will be rejected (i.e., MPCs of the two samples will differ) at 5% level of significance. We may thus conclude that the MPCs of two cases differ significantly.

In particular, if the pooled data are not given then

$$F^* = \frac{\sum e_1^2/(n_1-2)}{\sum e_2^2/(n_2-2)} \text{ with d.f} = \{(n_1-2),(n_2-2)\}$$

$$\text{Here } F^* = \frac{260.08/(35-2)}{226.6/(30-2)} \text{ with d.f } (33,28)$$

$$= \frac{7.8812}{8.0928} = 0.9739.$$

From the table value we see that $F_{0.05;33,\,28} = 1.84$ (approx.)

We may thus conclude that the null hypothesis will be accepted (as $F^* < F_{0.05;\,33,\,28}$) at 5% level of significance and hence there will be no difference in MPCs in the two cases.

2.18. Extension of Linear Regression Model to Non-linear Relationships

In the simple linear regression model we consider a linear relation between two variables X and Y in the form $Y_i = \alpha + \beta X_i + u_i$. But in many situations this may not be the case. In economics we observe non-linear relationships among the variables.

Some of the most common forms of non-linear relations used in economics are given below.

(i) Demand curve with unit elasticity : $D = f(P)$, or $D_i = \frac{\alpha}{P_i}$; where D represents quantity demanded and P denotes price.

(ii) Average cost curve : The traditional theory of 'U' shaped cost curve may be approximated by a polynomial of third degree in output :

$$C = C(q) \text{ or, } C_i = \alpha + \beta_1 q_i + \beta_2 q_i^2 + \beta_3 q_i^3 + u_i$$

where C represents cost and q represents the level of output.

$$\text{Now Average cost} = \frac{C_i}{q_i} = \frac{\alpha}{q_i} + \beta_1 + \beta_2 q_i + \beta_3 q_i^2 \text{, which is '}U\text{' shaped curve.}$$

(iii) Production function may be of the form : $Q = f(K, L)$ or $Q_i = AL_i^{\alpha} K_i^{\beta} u_i$ where Q = level of output, L = labour employed, K = capital employed, α and β are two parameters. This type of production function is called Cobb-Douglas production function.

(iv) The production function may be of the form :

$$Q_i = A\left[\delta K_i^{-\rho} + (1-\delta)L_i^{-\rho}\right]^{-\frac{1}{\rho}}.$$

This type of production function is called CES production function. The symbols have their usual meaning.

Now to estimate the parameters of the non linear functions first we have to convert the non linear function into a linear form and then we have to apply the OLS method in the usual manner.

In order to find out the rate of change (i.e., the slope) of the regressand (both linear and non linear but converted into linear form) with respect to the regressor as well as to find out the elasticity of the regressand with respect to the regressor we may consult the following table. The knowledge of these formulas will help us to compose the various models.

Model	Equation	Slope $\left(=\frac{dY}{dX}\right)$	Elasticity $=\left(\frac{X}{Y}\frac{dY}{dX}\right)$
Linear	$Y=\alpha+\beta X$	β	$\beta\cdot\left(\frac{X}{Y}\right)$*
Log-linear	$\log Y=\alpha+\beta\log X$	$\beta\cdot\left(\frac{Y}{X}\right)$	β
Log-linear	$\log Y=\alpha+\beta X$	$\beta\cdot(Y)$	$\beta\cdot(X)$*
Linear-log	$Y=\alpha+\beta\log X$	$\beta\cdot\left(\frac{1}{X}\right)$	$\beta\cdot\left(\frac{1}{Y}\right)$*
Reciprocal	$Y=\alpha+\beta\left(\frac{1}{x}\right)$	$-\beta\cdot\left(\frac{1}{X^2}\right)$	$-\beta\cdot\left(\frac{1}{XY}\right)$*
Log reciprocal	$\log Y=\alpha-\beta\left(\frac{1}{X}\right)$	$\beta\cdot\left(\frac{Y}{X^2}\right)$	$\beta\cdot\left(\frac{1}{X}\right)$*

Note : * indicates that the elasticity is variable, depending on the value taken by X or Y or both. When no X and Y values are specified, in practice, very often these elasticities are measured at the mean values of these variables, namely, $\overline{X}$ and $\overline{Y}$.

Example 2.10. Estimate the investment function $I=f(r)=\alpha(r)^{\beta}u$ on the basis of the following information :

$$n=11,\ \sum_{i=1}^{n}Y_i=12.2771,\ \sum_{i=1}^{n}X_i=16.6729$$

$$\sum_{i=1}^{n}X_i^2=27.9605,\ \sum_{i=1}^{n}X_iY_i=15.1222,$$

$$\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})=-3.4864,\ \sum_{i=1}^{n}(X_i-\overline{X})^2=2.6891,$$

$$\sum_{i=1}^{n}(Y_i-\overline{Y})^2=4.8566,$$

where $Y=\log I$, $X=\log r$.

Solution : The investment function is given by $I = \alpha(r)^{\beta} u$ which is a non-linear form, where α and β are the two parameters whose values are to be estimated by the OLS method. Taking log on both sides we get,

$$\log I = \log\alpha + \beta\log r + \log u \quad \text{or,} \quad I^* = \alpha^* + \beta r^* + u^*$$

where $\log I = I^*$, $\log\alpha = \alpha^*$, $\log r = r^*$ and $\log u = u^*$. This transformed function is linear in terms of logarithms.

The function $I^* = \alpha^* + \beta r^* + u^*$ is of the form $Y = \alpha + \beta X + u$ and hence we can apply the OLS method.

Now by the OLS method we can obtain the estimates of the parameters α^* and β.

Thus we have :

$$\text{(i)}\ \hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{-3.4864}{2.6891} = -1.12965$$

$$\begin{aligned}\text{and}\ \hat{\alpha}^* = \{\overline{Y} - \hat{\beta}\overline{X}\} &= \{1.1161 - (-1.2965) \times (1.5157)\}\\ &= [1.1161 + 1.9651]\\ &= 3.0812\end{aligned}$$

$$\left[\text{Here } \overline{Y} = \sum_{i=1}^{n} Y_i / n = \frac{12.27771}{11} = 1.1161 \text{ and } \overline{X} = \sum_{i=1}^{n} X_i / n = \frac{16.6729}{11} = 1.5157\right].$$

$$\text{(ii)}\ R^2 = \frac{\hat{\beta}^2\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = \frac{\hat{\beta}^2\sum_{i=1}^{n}x_i^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = \frac{\hat{\beta}^2\sum_{i=1}^{n}x_i^2}{\sum_{i=1}^{n}y_i^2}$$

$$= \frac{(-1.12965)^2 \times (2.6891)}{4.8566} = 0.70 \quad \therefore R^2 = 0.70$$

$$\text{(iii)}\ \hat{\sigma}_u^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum_{i=1}^{n}y_i^2 - \hat{\beta}^2\sum x_i^2}{n-2}$$

$$= \frac{4.8566 - (-1.12965)^2 \times 2.6891}{11-2} = \frac{4.8566 - 4.5201}{9} = \frac{0.3365}{9} = 0.0373.$$

$$\text{(iv)}\ Var(\hat{\beta}) = \frac{\hat{\sigma}^2 u}{\sum x_i^2} = \frac{0.0373}{2.6891} = 0.01387$$

$$\therefore SE(\hat{\beta}) = \sqrt{0.01387} = 0.1177.$$

(v) The regression results can now be written as follows :

$\hat{I}^* = \hat{\alpha}^* + \hat{\beta} r^*$

or, $\log I = 3.0812 - 1.12965 \log r \qquad R^2 = 0.70$
$\qquad\qquad\qquad\qquad (0.1177)$

or, $\hat{I} = \alpha \cdot (r)^{-1.2965}$ where α = antilog of 3.0812.

The results show that the constant interest elasticity is –1.2965. This means that the demand for investment is interest elastic. This result is consistent with the existing theory.

2.19. Problem of Prediction / Forecasting Relating to a Two-Variable Linear Regression Model

Usually we do not differentiate between prediction and forecasting. We use one or the other interchangeably. But these two terms are not identical. Prediction simply means an estimation of any event happening (in the past, present or future). On the other hand, forecasting is always associated with a time dimension in the future i.e., estimation for some specific future duration or over a period of time. All forecasts are predictions, but not all predictions are forecasts, as when we use regression to explain the relationship between two variables 'forecast' implies time series and future while prediction does not. When we are interested to predict or forecast about future, we call the regression analysis as **historical regression.** With the help of regression analysis we can forecast about the future value on the basis of past and present information of the said variables (X and Y). In the context of forecasting we may also distinguish between *ex-ante forecast* and *ex-post forecast.* Ex-ante forecast is a forecast that uses information available at the time of forecast, whereas expost forecast is a forecast that uses information beyond the time at which the forecast is made.

Let us define a classical linear regression model, given by $Y_i = \alpha + \beta X_i + u_i$ for i = 1, 2, ..., n with the help of the pairs of observations (X_1, Y_1), (X_2, Y_2), ... (X_n, Y_n). We estimate the relationship by the method of least squares. The estimated relationship is $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$. In case of time series data we write the regression equation as $Y_t = \alpha + BX_t + u_t$.

Now for some value of X, (the independent variable), which is not in the sample, we may like to estimate the value of Y (the dependent variable).

The process of finding the value of the dependent variable from the estimated relationship for the known value of the independent variable, not in the sample, is called "Prediction".

Let us suppose that X_0 is the value of the independent variable, not in the sample and we have to predict the value of Y when $X = X_0$. There are two types of prediction: (i) Point prediction, (ii) Interval prediction.

2.19.1. Point Prediction

When prediction is done in terms of a single value of the dependent variable, then it is called point prediction. We simply put $X = X_0$ in the estimated relationship and we get, $\hat{Y} = \hat{\alpha} + \hat{\beta} X_0 = \hat{Y}_0$.

Now, Y_0 is the true value of the dependent variable Y when prediction is made and it is given by,

$Y_0 = \alpha + \beta X_0 + u_0$ where u_0 is the corresponding value of the disturbance term.

So, $E(Y_0) = E(\alpha) + \beta E(X_0) + E(u_0)$

$= \alpha + \beta X_0$, $[\therefore E(X_0) = X_0, E(u_0) = 0]$

Let us define the prediction error by $e_0 = Y_0 - \hat{Y}_0$ when we want to predict Y_0

and $e_0 = E(Y_0) - \hat{Y}_0$ is the prediction error when we want to predict $E(Y_0)$.

Mean of the Predictor :

When we want to predict Y_0, then $e_0 = Y_0 - \hat{Y}_0$

Now $e_0 = Y_0 - \hat{Y}_0 = \alpha + \beta X_0 + u_0 - \hat{\alpha} - \hat{\beta} X_0$

or, $e_0 = u_0 - (\hat{\alpha} - \alpha) - (\hat{\beta} - \beta) X_0$

or, $E(e_0) = E(u_0) - E(\hat{\alpha} - \alpha) - E(\hat{\beta} - \beta) X_0 = 0$

where $E(u_0) = 0,\ E(\hat{\alpha}) = \alpha,\ E(\hat{\beta}) = \beta$.

Hence $E(Y_0 - \hat{Y}_0) = 0$ i.e., $E(e_0) = 0$.

or, $Y_0 - E(\hat{Y}_0) = 0$ or, $E(\hat{Y}_0) = Y_0$

So, $\hat{Y}_0$, the OLS point predictor of Y_0 is unbiased.

Variance of the predictor : Variance of the predictor is given by,

$$Var(\hat{Y}_0) = E\left[\hat{Y}_0 - E(\hat{Y}_0)\right]^2 = E\left[\hat{Y}_0 - Y_0\right]^2 \left[\because E(\hat{Y}_0) = Y_0\right]$$

$$= E[-e_0]^2 = E(e_0)^2 \quad [\because e_0 = Y_0 - \hat{Y}_0]$$

$$\text{Now } E(e_0^2) = E\left[u_0 - (\hat{\alpha} - \alpha) - (\hat{\beta} - \beta) X_0\right]^2$$

$$= E[u_0^2 + (\hat{\alpha} - \alpha)^2 + (\hat{\beta} - \beta)^2 X_0^2 - 2(\hat{\alpha} - \alpha) u_0$$

$$-2(\hat{\beta} - \beta) X_0 u_0 + 2(\hat{\alpha} - \alpha)(\hat{\beta} - \beta) X_0]$$

$$= E(u_0^2) + E(\hat{\alpha} - \alpha)^2 + E(\hat{\beta} - \beta)^2 X_0^2 - 2E(\hat{\alpha} - \alpha) u_0$$

$$-2X_0 E(\hat{\beta} - \beta) u_0 + 2X_0 E(\hat{\alpha} - \alpha)(\hat{\beta} - \beta)$$

$$= \sigma_u^2 + \text{var}(\hat{\alpha}) + X_0^2 \text{ var}(\hat{\beta}) - 2\text{cov}(\hat{\alpha}, u_0)$$

$$-2X_0 \text{ cov}(\hat{\beta}, u_0) + 2X_0 \text{ cov}(\hat{\alpha}, \hat{\beta})$$

We know that, $\text{var}(\hat{\alpha}) = \sigma_u^2 \left[\dfrac{1}{n} + \dfrac{\overline{X}^2}{\sum_{i=1}^{n} x_i^2}\right]$ and $\text{cov}(\hat{\alpha}, u_0) = \text{cov}(\hat{\beta}, u_0) = 0$ as the OLS estimators of the parameters are independent of the disturbance term.

Now $\text{cov}(\hat{\alpha},\hat{\beta}) = E\left[(\hat{\alpha}-\alpha)(\hat{\beta}-\beta)\right]$

$= E\left[(\bar{Y}-\hat{\beta}\bar{X}-\bar{Y}+\beta\bar{X})(\hat{\beta}-\beta)\right] = E\left[-\bar{X}(\hat{\beta}-\beta)(\hat{\beta}-\beta)\right]$

$= -\bar{X}E(\hat{\beta}-\beta)^2 = -\bar{X}\,\text{var}(\hat{\beta})$

So, $\text{var}(\hat{Y}_0) = E(e_0^2)$

$= \sigma_u^2 + \text{var}(\hat{\alpha}) + X_0^2\,\text{var}(\hat{\beta}) - 2X_0\bar{X}\,\text{var}(\hat{\beta})$

$\therefore \text{var}(\hat{Y}_0) = \sigma_u^2 + \text{var}(\hat{\alpha}) + X_0^2\,\text{var}(\hat{\beta}) - 2X_0\bar{X}\,\text{var}(\hat{\beta}) + \bar{X}^2\,\text{var}(\hat{\beta}) - \bar{X}^2\,\text{var}(\hat{\beta})$

$= \sigma_u^2 + \text{var}(\hat{\alpha}) + (X_0-\bar{X})^2\,\text{var}(\hat{\beta}) - \bar{X}^2\,\text{var}(\hat{\beta})$

$$\left[\text{Since, } \text{var}(\hat{\alpha}) = \sigma_u^2\left(\frac{1}{n}+\frac{\bar{X}^2}{\sum_{i=1}^{n}x_i^2}\right) = \frac{\sigma_u^2}{n}+\bar{X}^2\frac{\sigma_u^2}{\sum_{i=1}^{n}x_i^2} = \frac{\sigma_u^2}{n}+\bar{X}^2\,\text{var}(\hat{\beta})\right]$$

$\therefore \text{var}(\hat{Y}_0) = \sigma_u^2 + \frac{\sigma_u^2}{n} + \bar{X}^2\,\text{var}(\hat{\beta}) + (X_0-\bar{X})^2\,\text{var}(\hat{\beta}) - \bar{X}^2\,\text{var}(\hat{\beta})$

or, $\text{var}(\hat{Y}_0) = E(e_0^2) = \sigma_u^2\left(1+\frac{1}{n}\right) + (X_0-\bar{X})^2\,\text{var}(\hat{\beta})$

Except $\text{var}(\hat{\beta})$ all the terms are constant and positive.

So, $\text{var}(\hat{Y}_0)$ is minimum when $\text{var}(\hat{\beta})$ is minimum.

We know that, $\text{var}(\hat{\beta})$ is minimum when $\hat{\beta}$ is the OLS estimator of β. Hence, $\text{var}(\hat{Y}_0)$ is minimum when $\hat{Y}_0$ is the OLS point-predictor of Y_0. This is the BLUE property of OLS point predictor. It means that OLS point predictor of Y_0 i.e., $\hat{Y}_0$ is the best linear unbiased predictor of Y_0.

We now consider another case where we want to make a **point prediction of $E(Y_0)$**. Here prediction error is defined as $E(Y_0) - \hat{Y}_0 = e_0$. We know that, $Y_0 = \alpha + \beta X_0 + u_0$.

$\therefore E(Y_0) = \alpha + \beta X_0$ because $E(u_0) = 0$

$\therefore e_0 = \alpha + \beta X_0 - \hat{\alpha} - \hat{\beta}X_0$

$= (\alpha-\hat{\alpha}) + (\beta-\hat{\beta})X_0 = -(\hat{\alpha}-\alpha) - (\hat{\beta}-\beta)X_0$

$\therefore e_0^2 = (\hat{\alpha}-\alpha)^2 + (\hat{\beta}-\beta)^2X_0^2 + 2X_0(\hat{\alpha}-\alpha)(\hat{\beta}-\beta)$

or, $E(e_0^2) = E(\hat{\alpha}-\alpha)^2 + X_0^2E(\hat{\beta}-\beta)^2 + 2X_0E(\hat{\alpha}-\alpha)(\hat{\beta}-\beta)$

$\therefore E(e_0^2) = E\left[\hat{Y}_0 - E(Y_0)\right]^2 = \text{var}(\hat{\alpha}) + X_0^2\,\text{var}(\hat{\beta}) + 2X_0\,\text{cov}(\hat{\alpha},\hat{\beta})$

$$\left[\text{Since } \operatorname{var}(\hat{\alpha}) = \sigma_u^2\left(\frac{1}{n}+\frac{\bar{X}^2}{\sum_{i=1}^{n} x_i^2}\right) = \frac{\sigma_u^2}{n}+\bar{X}^2 \cdot \frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2} = \frac{\sigma_u^2}{n}+\bar{X}^2 \operatorname{var}(\hat{\beta})\right]$$

and $\operatorname{cov}(\hat{\alpha},\hat{\beta}) = E[(\hat{\alpha}-\alpha)(\hat{\beta}-\beta)]$

$= E\left[(\bar{Y}-\hat{\beta}\bar{X}-\bar{Y}+\beta\bar{X})(\hat{\beta}-\beta)\right]$ $[\because \hat{\alpha} = \bar{Y}-\hat{\beta}\bar{X}$ and $\alpha = \bar{Y} = \beta\bar{X}]$

$= E[-\bar{X}(\hat{\beta}-\beta)(\hat{\beta}-\beta)] = -\bar{X}E(\hat{\beta}-\beta)^2 = -\bar{X}\operatorname{var}(\hat{\beta})]$

Now putting the values of $\operatorname{var}(\hat{\alpha})$ and $\operatorname{cov}(\hat{\alpha},\hat{\beta})$ in the above expression we get

$E\left[\hat{Y}_0 - E(Y_0)\right]^2 = \operatorname{var}(\hat{Y}_0) = \operatorname{var}(e_0) = E(e_0^2)$

Now $\operatorname{var}(e_0) = E\left[\hat{Y}_0 - E(Y_0)\right]^2$

$= \frac{\sigma_u^2}{n}+\bar{X}^2\operatorname{var}(\hat{\beta})+X_0^2\operatorname{var}(\hat{\beta})-2X_0\bar{X}\operatorname{var}(\hat{\beta}) = \frac{\sigma_u^2}{n}+(X_0-\bar{X})^2\operatorname{var}(\hat{\beta})$

var (e_0) is minimum when $\operatorname{var}(\hat{\beta})$ is minimum. Now $\operatorname{var}(\hat{\beta})$ is minimum where $\hat{\beta}$ is the OLS estimator of β. So $\operatorname{var}(e_0)$ is minimum when $\hat{Y}_0$ is the OLS point predictor of $E(Y_0)$. This is the BLUE property of the OLS point predictor $E(Y_0)$.

2.19.2. Test of Significance of Predictor and Interval Prediction

Case 1 : We want to test the null hypothesis $H_0 : Y_0 = A$ (some specified value against the alternative hypothesis $H_1 : Y_0 \neq A$ or, $H_1 : Y_0 > A$ or, $H_1 : Y_0 < A$. We use $\hat{Y}_0$ as the appropriate statistic of Y_0 because $\hat{Y}_0$ is the BLUE predictor of Y_0. Now $\hat{Y}_0 = \hat{\alpha}+\hat{\beta}X_0$. Since $\hat{Y}_0$ is a linear function of $\hat{\alpha}$ and $\hat{\beta}$. $\hat{\alpha}$ and $\hat{\beta}$ are normally distributed. So $\hat{Y}_0$ is also normally distributed.

Since $E(e_0) = E\left[\hat{Y}_0 - Y_0\right] = 0, \ \therefore E(\hat{Y}_0) = Y_0$

and $\operatorname{var}(\hat{Y}_0) = E\left[\hat{Y}_0 - Y_0\right]^2 = E(e_0^2)$

$= \sigma_u^2 + \operatorname{var}(\hat{\alpha}) + \operatorname{var}(\hat{\beta})X_0^2 + 2X_0\operatorname{cov}(\hat{\alpha},\hat{\beta})$

$$= \sigma_u^2 + \sigma_u^2\left[\frac{1}{n}+\frac{\bar{X}^2}{\sum_{i=1}^{n} x_i^2}\right] + X_0^2\frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2} - 2X_0\bar{X}\frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2}, \text{ where } \operatorname{cov}(\hat{\alpha},\hat{\beta}) = -\bar{X}\operatorname{var}(\hat{\beta})$$

and $\operatorname{var}(\hat{\beta}) = \frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2}$

$$= \sigma_u^2\left[1+\frac{1}{n}+\frac{1}{\sum\limits_{i=1}^{n} x_i^2}(\bar{X}^2+X_0^2-2X_0\bar{X})\right] = \sigma_u^2\left[1+\frac{1}{n}+\frac{1}{\sum\limits_{i=1}^{n} x_i^2}(\bar{X}-X_0)^2\right]$$

$$\therefore \operatorname{var}(\hat{Y}_0) = \sigma_u^2\left[1+\frac{1}{n}+\frac{1}{\sum\limits_{i=1}^{n} x_i^2}(\bar{X}-X_0)^2\right]$$

This means that $\hat{Y}_0$ is normally distributed with mean Y_0 and variance

$\sigma_u^2\left[1+\frac{1}{n}+\frac{(\bar{X}-X_0)^2}{\sum\limits_{i=1}^{n} x_i^2}\right]$. So, if σ_u^2 is unknown, it is to be replaced by its unbiased

estimator $\sum\limits_{i=1}^{n} e_i^2/(n-2)$. The appropriate test statistic will be given by,

$$\frac{\hat{Y}_0 - Y_0}{\sqrt{\left(\frac{\sum\limits_{i=1}^{n} e_i^2}{n-2}\right)\left[1+\frac{1}{n}+\frac{(\bar{X}-X_0)^2}{\sum\limits_{i=1}^{n} x_i^2}\right]}} \sim t_{n-2}.$$

It follows a 't' distribution with $(n-2)$ degrees of freedom.

Nature of the test : If the alternative hypothesis is $H_1 : Y_0 \neq A$, then the null hypothesis will the accepted at 5% level of significance if $-t_{0.025}, n-2 \leq t \leq t_{0.025}, n-2$ and will be rejected otherwise.

If the alternative hypothesis is $H_1 : Y_0 > A$ then $H_0 : Y_0 = A$ will be accepted at 5% level of significance if t (observed) $\leq t_{0.05}, n-2$ (table) and will be rejected otherwise.

If the alternative hypothesis is $H_1 : Y_0 < A$, then $H_0 : Y_0 = A$ will be accepted at 5% level of significance if

t (observed) $\geq -t_{0.05}, n-2$ (table), and will be rejected otherwise.

The rejection of the null hypothesis on the basis of the sample data implies the significance of Y_0.

It should be noted that the same procedure can be used for 1% level of significance.

It can be seen that 100 (1 – α)% confidence interval of Y_0 would be

$$\hat{Y}_0 \pm t_{\alpha/2, n-2} \cdot \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}\left(1+\frac{1}{n}+\frac{(\bar{X}-X_0)^2}{\sum x_i^2}\right)}$$ where α usually takes the value 0.05 or 0.01.

Case 2. We want to test the null hypothesis $H_0 : E(Y_0) = A$ against the alternative $H_1 : E(Y_0) \neq A$ or, $H_1 : E(Y_0) > A$ or, $H_1 : E(Y_0) < A$.

Here we take $\hat{Y}_0$ as the statistic of $E(Y_0)$ because $\hat{Y}_0$ is the BLUE predictor of $E(Y_0)$.

Here also $\hat{Y}_0$ is normally distributed with mean $E(Y_0)$ and variance

$$\operatorname{var}(\hat{Y}_0) = \frac{\sigma_u^2}{n} + \left(X_0 - \bar{X}\right)^2 \operatorname{var}(\hat{\beta})$$

$$= \frac{\sigma_u^2}{n} + \left(X_0 - \bar{X}\right)^2 \cdot \frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2} = \sigma_u^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} x_i^2}\right]$$

So, $\hat{Y}_0$ is normally distributed with mean $E(Y_0)$ and variance $\sigma_u^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} x_i^2}\right]$

If σ_u^2 is unknown, it is to be replaced by its unbiased estimator $\sum_{i=1}^{n} e_i^2 \Big/ (n-2)$. The appropriate test statistic under $H_0 : E(Y_0) = A$ would be

$$\frac{\hat{Y}_0 - E(Y_0)}{\sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}\left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} x_i^2}\right)}} \sim t_{n-2}$$

which follows a t-distribution with $(n-2)$ degrees of freedom.

Nature of the test :

If $H_0 : E(Y_0) = A$ is tested against the alternative $H_1 : E(Y_0) \neq A$, then $H_0 : E(Y_0) = A$ will be accepted at 5% level of significance if $-t_{0.025, n-2} \leq t \text{ (observed)} \leq t_{0.025, n-2}$ and will be rejected otherwise.

If the alternative hypothesis is $H_1 : E(Y_0) > A$, then $H_0 : E(Y_0) = A$ will be accepted at 5% level of significance if t (observed) $\leq t_{0.05}, n-2$ (table) and will be rejected otherwise.

If the alternative hypothesis is $H_1 : E(Y_0) < A$, then $H_0 : E(Y_0) = A$ will be accepted at 5% level of significane if t (observed) $\geq -t_{0.05}, n-2$ (table) and will be rejected otherwise. The same test procedure is applicable at 1% level of significance.

It can be seen that 100 (1 – α)% confidence interval of $E(Y_0)$ would be

$$\hat{Y}_0 \pm t_{\alpha/2, n-2} \cdot \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}\left(\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum x_i^2}\right)}$$ where α usually takes the value 0.05 or 0.01.

Example : 2.11. Following example 2.5,

(i) find out the point predictor of Y_i when $X_i = 10$.

(ii) It is claimed that when $X_i = 10$, $Y_i = 165$. Do you think that it is justified ?

(iii) It is claimed that when $X_i = 10$, $E(Y_i) = 155$. Do you think that the claim is justified?

Solution :

(i) We have to find out the point predictor of Y_i when $X_i = 10$. We know that the point predictor of Y_i is,

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i \qquad \text{(where } \hat{\alpha} = -3.505,\ \hat{\beta} = 0.494\text{. See Ex. 2.5)}$$

$$= -3.505 + 0.494 \times 10$$

$$= -3.505 + 4.94 = 1.435$$

$$\therefore \hat{Y}_i = 1.435$$

So, point predictor of Y_i is $\hat{Y}_i = 1.435$ when $X_i = 10$.

(ii) We have to examine whether $Y_i = 165$ when $X_i = 10$ is justified or not. We have to test the null hypothesis $H_0 : Y_0 = 165$, against the alternative $H_1 : Y_0 \neq 165$.

The appropriate test statistic is given by, $$\frac{\hat{Y}_0 - Y_0}{\sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}\left[1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n} x_i^2}\right]}} \sim t_{n-2}.$$

Since $\hat{\alpha} = -3.505$, $\hat{\beta} = 0.494$ and $X_0 = 10$,

then $\hat{Y}_0 = \hat{\alpha} + \hat{\beta}X_0 = -3.505 + 0.494 \times 10 = 1.435$

$$\therefore \hat{Y}_0 - Y_0 = 1.435 - 165.000 = -163.565$$

and $\sqrt{\dfrac{\sum_{i=1}^{n} e_i^2}{n-2}\left[1+\dfrac{1}{n}+\dfrac{(\bar{X}-X_0)^2}{\sum_{i=1}^{n} x_i^2}\right]}$

$$=\sqrt{1.908\times\left[1+\frac{1}{20}+\frac{(9.31-10)^2}{215.4}\right]}$$

$$=\sqrt{1.908\times[1+0.05+0.00221]}$$

$$=\sqrt{1.908\times 1.05221}=\sqrt{2.0076167}$$

$$=1.4169$$

Since $\dfrac{\sum_{i=1}^{n} e_i^2}{n-2}=\dfrac{34.34}{20-2}$

$$=\frac{34.34}{18}=1.908,$$

$$\bar{X}=9.31$$

and $\sum_{i=1}^{n} x_i^2=215.4$

Now t (observed) $=\dfrac{\hat{Y}_0-Y_0}{\sqrt{\dfrac{\sum_{i=1}^{n} e_i^2}{n-2}\left[1+\dfrac{1}{n}+\dfrac{(\bar{X}-X_0)^2}{\sum_{i=1}^{n} x_i^2}\right]}}$

$$=\frac{-163.565}{1.4169}=-115.438$$

$\therefore t=-115.438$

Here we see that t (observed) = -115.438 which does not lie in the interval $-t_{0.025}$, 18 and $t_{0.025}$, 18 i.e., in the interval -2.101 and 2.101 and hence the null hypothesis $H_0 : Y_0 = 165$ is rejected for the given sample at 5% level of significance. So, $Y_0 = 165$ is not justified when $X_0 = 10$.

(iii) We have to examine whether $E(Y_i) = 155$ when $X_i = 10$.

We have to test the null hypothesis $H_0 : E(Y_0) = 155$, against the alternative $H_1 : E(Y_0) \neq 155$.

The statement $E(Y_i) \neq 155$, when $X_i = 10$ will be justified if the null hypothesis is $H_0 : E(Y_0) = 155$ is rejected.

The appropriate test statistic would be :

$$t=\frac{\hat{Y}_0-E(Y_0)}{\sqrt{\dfrac{\sum_{i=1}^{n} e_i^2}{n-2}\left(\dfrac{1}{n}+\dfrac{(X_0-\bar{X})^2}{\sum_{i=1}^{n} x_i^2}\right)}}\sim t_{n-2}$$

Here $\hat{Y}_0 = \hat{\alpha} + \hat{\beta}\bar{X} = 1.435,\ E(Y_0) = 155.$

$\therefore \hat{Y}_0 - E(Y_0) = 1.435 - 155.000 = -153.565$

Since $\sum_{i=n}^{n} e_i^2 \Big/ (n-2) = \frac{34.34}{20-2} = \frac{34.34}{18} = 1.908$

$n = 20,\ X_0 = 10,\ \bar{X} = 9.31,\ \sum_{i=1}^{n} x_i^2 = 215.4$

$$\therefore \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}\left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} x_i^2}\right)}$$

$$= \sqrt{1.908\left[\frac{1}{20} + \frac{(10-9.31)^2}{215.4}\right]}$$

$$= \sqrt{1.908[0.05 + 0.00221]} = \sqrt{1.908 \times 0.05221} = \sqrt{0.09961} = 0.3156.$$

$$\text{Now } t \text{ (observed)} = \frac{\hat{Y}_0 - E(Y_0)}{\sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}\left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} x_i^2}\right)}} = \frac{-153.565}{0.3156} = -486.581$$

t (observed) = – 486.581.

Now on the basis of the given sample the null hypothesis $H_0 : E(Y_0) = 155$ will be accepted at 5% level of significance if $-t_{0.025,n-2} \le t \le t_{0.025,n-2}$ and will be rejected otherwise.

Here $t_{0.025,n-2} = t_{0.025,18} = 2.101$.

So, the observed $t = -486.581$ does not lie in the interval –2.101 and 2.102 and hence the null hypothesis will be rejected. So, $E(Y_0) = 155$ when $X_0 = 10$ is not justified.

Example 2.12. Consider the following regression model $Y_i = \alpha + \beta X_i + u_i$ where u_i is normally distributed with mean zero and variance σ_u^2 (unknown). We have the following data :

X	:	2	3	1	5	9
Y	:	4	7	3	9	7

(i) Estimate α and β.

(ii) Test whether $\hat{\alpha}$ and $\hat{\beta}$ are significant or not at 5% level of significance.

(iii) Calculate R^2.

(iv) Find out the point predictor of Y when $X = 10$.

Solution : We have to estimate the regression parameters α and β. Let $\hat{\alpha}$ and $\hat{\beta}$ be the OLS estimators of α and β. We know that,

$$\hat{\beta} = \sum_{i=1}^{n} x_i y_i \Big/ \sum_{i=1}^{n} x_i^2 \text{ and } \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}.$$

Calculations for $\hat{\alpha}$, $\hat{\beta}$ and R^2

X_i	Y_i	X_iY_i	X_i^2	x_i $= X_i - \bar{X}$	y_i $Y_i - \bar{Y}$	x_i^2	y_i^2	x_iy_i
2	4	8	4	–2	–2	4	4	4
3	7	21	9	–1	1	1	1	–1
$n = 5$, 1	3	3	1	–3	–3	9	9	9
5	9	45	25	1	3	1	9	3
9	7	63	81	5	1	25	1	5
$\sum_{i=1}^{n} X_i = 20$	$\sum_{i=1}^{n} Y_i = 30$	$\sum_{i=1}^{n} X_iY_i = 140$	$\sum_{i=1}^{n} X_i^2 = 120$	$\sum_{i=1}^{n} x_i = 0$	$\sum_{i=1}^{n} y_i = 0$	$\sum_{i=1}^{n} x_i^2 = 40$	$\sum_{i=1}^{n} y_i^2 = 24$	$\sum_{i=1}^{n} x_iy_i = 20$

Now $\bar{X} = \sum_{i=1}^{n} X_i \big/ n = \frac{20}{5} = 4$ and $\bar{Y} = \sum_{i=1}^{n} Y_i \big/ n = \frac{30}{5} = 6$

Now $\hat{\beta} = \sum_{i=1}^{n} x_i y_i \Big/ \sum_{i=1}^{n} x_i^2 = \frac{20}{40} = 0.5$ and $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 6 - 0.5 \times 4 = 6 - 2 = 4$

$\therefore \hat{\alpha} = 4, \hat{\beta} = 0.5$

Thus the OLS estimators of α and β are $\hat{\alpha} = 4$ and $\hat{\beta} = 0.5$.

The estimated regression line is $\hat{Y} = \hat{\alpha} + \hat{\beta}X$ or, $\hat{Y} = 4 + 0.5X$

Now we have to find out $Var(\hat{\alpha})$ and $Var(\hat{\beta})$. We know that,

$$Var(\hat{\alpha}) = \sigma_u^2 \left(\frac{\sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} x_i^2} \right) \text{ and } Var(\hat{\beta}) = \sigma_u^2 \Big/ \sum_{i=1}^{n} x_i^2 .$$

But here σ_u^2 is not known and hence it is to be replaced by its unbiased estimator $\hat{\sigma}_u^2 = \sum_{i=1}^{n} e_i^2 \Big/ (n-2)$.

Now $\sum_{i=1}^{n} e_i^2 \Big/ (n-2) = \dfrac{\sum_{i=1}^{n} y_i^2 - \hat{\beta}^2 \sum_{i=1}^{n} x_i^2}{n-2}$

$$= \frac{24-(0.5)^2 \times 40}{5-2} = \frac{24-10}{3} = \frac{14}{3} = 4.67$$

$\therefore \hat{\sigma}_u^2 = 4.67.$

Now $Var(\hat{\alpha}) = \hat{\sigma}_u^2 \left(\dfrac{\sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} x_i^2} \right) = 4.67 \times \dfrac{120}{5 \times 40} = 4.67 \times \dfrac{120}{200} = 2.802$

$\therefore Var(\hat{\alpha}) = 2.802$ and $SE(\hat{\alpha}) = \sqrt{Var(\hat{\alpha})} = \sqrt{2.802} = 1.674$

Similarly, $Var(\hat{\beta}) = \hat{\sigma}_u^2 \Big/ \sum_{i=1}^{n} x_i^2 = \dfrac{4.67}{40} = 0.11675$

$\therefore Var(\hat{\beta}) = 0.11675$ and $SE(\hat{\beta}) = \sqrt{Var(\hat{\beta})} = \sqrt{0.11675} = 0.3416.$

(ii) **Test for** $\hat{\alpha}$ **and** $\hat{\beta}$ **:**

(a) **Test for** $\hat{\beta}$: We have to test the null hypothesis $H_0 : \hat{\beta} = 0$ against the alternative $H_1 : \hat{\beta} \neq 0$. The appropriate test statistic would be :

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})} \sim t_{n-2}$$

The null hypothesis will be accepted at 5% level of significance if $-t_{0.025,n-2} \leq t \leq t_{0.025,n-2}$ and will be rejected otherwise.

Here we see that,

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})} = \frac{\hat{\beta}}{\sqrt{\left(\sum_{i=1}^{n} e_i^2 \Big/ (n-2)\right) \Big/ \sum_{i=1}^{n} x_i^2}} = \frac{0.5}{0.3416} = 1.4637$$

$\therefore t$ (observed) = 1.4637.

But from table value : $t_{0.025,n-2} = t_{0.025,\,5-2} = t_{0.025,\,3} = 3.182$

Here we see that the t (observed) = 1.4637 lies in the interval –3.182 and 3.182 and hence the null hypothesis is accepted at 5% level of significance.

So, $\hat{\beta}$ is insignificant at 5% level ; significant only when the null hypothesis is rejected.

(b) **Test for** $\hat{\alpha}$: If we test the null hypothesis $H_0 : \hat{\alpha} = 0$ against the alternative $H_1 : \hat{\alpha} \neq 0$ then the null hypothesis will be accepted at 5% level of significance if $-t_{0.025,n-2} \leq t \leq t_{0.025,n-2}$ and will be rejected otherwise. Here test statistic would be $t = \dfrac{\hat{\alpha}}{SE(\hat{\alpha})} \sim t_{n-2}$.

From table values $t_{0.025,\ n-2} = t_{0.025,\ 3} = 3.182$ (as $n = 5$)

$$\text{and } t \text{ (observed)} = \frac{\hat{\alpha}}{SE(\hat{\alpha})} = \frac{\hat{\alpha}}{\sqrt{\dfrac{\sum_{i=1}^{n} e_i^2}{n-2}\left(\dfrac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2}\right)}} = \frac{4}{1.674} = 2.3895$$

Thus we see that t (observed) $= 2.3895$ lies in the interval $-t_{0.025,\ n-2}$ and $t_{0.025,\ n-2}$ i.e., -3.182 and 3.182 and hence the null hypothesis is accepted at 5% level of significance. This means that $\hat{\alpha}$ is also insignificant at 5% level of significance.

(iii) Now we have to calculate the value of R^2.

$$\text{Since we know that, } R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\hat{\beta}^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2} = \frac{(0.5)^2 \cdot 40}{24} = \frac{10}{24} = 0.4167 \approx 0.42$$

$$\therefore R^2 = 0.42 = \frac{42}{100} = \frac{\text{Explained variation}}{\text{Total variation}}.$$

This suggests that 42 percent of the variations in the sample observations of Y can be attributed to the variations of the fitted value of Y, i.e., $\hat{y}$ or we can say that our regression line fits the given data not very well.

From the above results we can write our regression results as follows :

$$Y_i = 4 + 0.5X_i,\ R^2 = 0.42$$

SE : (1.674) (0.3416)

Alternatively,

$$Y_i = 4 + 0.5X_i,\ R^2 = 0.42$$

t ratios : (2.3895) (1.4637)

(iv) We have to find out the point predictor of Y when $X = 10$.

The point predictor of Y is given by $\hat{Y} = \hat{\alpha} + \hat{\beta}X$

$= 4 + 0.5 \times 10 = 4 + 5 = 9$

$\therefore$ Point predictor $\hat{y} = 9$ when $X = 10$.

Example 2.13. Following the data given in Example 2.1,

(i) estimate the regression parameters [assuming a linear regression equation of the form $Y_i = \alpha + \beta X_i + u_i$, where $u_i \sim N\ (0,\ \sigma_u^2$ (unknown)].

(ii) Calculate R^2

(iii) Forecast about the value of sales revenue if advertising expenditures are increased to ₹ 600.

(iv) Find out 95% confidence interval of predicted value of Y when X = ₹ 600.

(v) Find out 95% confidence interval of the expected predicted value of Y when X = ₹ 600.

Solution : **Table for calculation**

Month	X_i	Y_i	$x_i = X_i - \bar{X}$	$y_i = Y_i - \bar{Y}$	x_iy_i	x_i^2	y_i^2	$e_i = Y_i - \hat{Y}_i$	e_i^2
1	1	3	− 2	− 1.6	3.2	4	2.56	0.8	0.64
2	2	4	− 1	− 0.6	0.6	1	0.36	0.6	0.36
3	3	2	0	− 2.6	0	0	6.76	− 2.6	6.76
4	4	6	1	1.4	1.4	1	1.96	0.2	0.04
5	5	8	2	3.4	6.8	4	11.56	1.0	1.00
Total	$\Sigma X_i = 15$	$\Sigma Y_i = 23$	$\Sigma x_i = 0$	$\Sigma y_i = 0$	$\Sigma x_iy_i = 12$	$\Sigma x_i^2 = 10$	$\Sigma y_i^2 = 23.20$	$\Sigma e_i = 0$	$\Sigma e_i^2 = 8.8$

Here $n = 5$ as we have data for five months.

Now $\bar{X} = \dfrac{\sum X_i}{n} = \dfrac{15}{5} = 3$, $\bar{Y} = \dfrac{\sum Y_i}{n} = \dfrac{23}{5} = 4.6$

(i) We have to estimate the regression parameters α and β. Let $\hat{\alpha}$ and $\hat{\beta}$ be the OLS estimators (predictors here) of α and β.

We know that $\hat{\beta} = \dfrac{\sum x_i y_i}{\sum x_i^2}$ and $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$

$\therefore \hat{\beta} = \dfrac{12}{10} = 1.2$ and $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 4.6 - 1.2 \times 3 = 4.6 - 3.6 = 1$

Therefore the estimated (predicted) sample regression equation is $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i = 1.0 + 1.2X_i$

In the table $e_i = Y_i - \hat{Y}_i$ can be obtained for different values of X_i. Since $e_i = Y_i - 1.0 - 1.2X_i$.

When $X_i = 1$ and $Y_i = 3$, $e_i = 3 - 1.0 - 1.2 \times 1 = 3.0 - 2.2 = 0.8$
$X_i = 2$ and $Y_i = 4$, $e_i = 4 - 1.0 - 1.2 \times 2 = 4 - 1 - 2.4 = 0.6$
and in this way other e_i values are calculated.

(ii) We know that $R^2 = \dfrac{\text{ESS}}{\text{TSS}} = \dfrac{\hat{\beta}^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2} = \dfrac{(1.2)^2 \times 10}{23.20} = \dfrac{14.4}{23.20} = 0.620$

Since $\sum_{i=1}^{n} y_i^2 = \hat{\beta}^2 \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} e_i^2$ i.e., TSS = ESS + RSS $\therefore R^2 = 0.620$

(iii) Since predicted regression equation is given by $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i = 1.0 + 1.2X_i$

So, when $X = X_0 = 6$, then $\hat{Y}_0 = 1.0 + 1.2X_0 = 1.0 + 1.2 \times 6$

$\therefore \hat{Y}_0 = 8.2$

This means that if advertisement expenditures are increased to ₹ 600, sales revenue becomes ₹ 8200.

(iv) We have to find out 95% confidence interval of the predicted value of Y [$\hat{Y}_0$] when X = ₹ 600.

We know that at 100 $(1-\alpha)$% [$\alpha = 0.05$ here] confidence interval of the true forecast (prediction) is

$$\hat{Y}_0 \pm t_{\alpha/2, n-2}\sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}\left[1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum x_i^2}\right]}$$

Here we have $\hat{Y}_0 = 8.2$ when $X = X_0 = 6$

and $\sum_{i=1}^{n} e_i^2 = 8.8$, $n = 5$, $\sum_{i=1}^{n} x_i^2 = 10$, $\bar{X} = 3$, $\frac{\sum_{i=1}^{n} e_i^2}{(n-2)} = \frac{8.8}{3} = 2.93$

and $t_{\alpha/2, n-2} = t_{0.025, 3} = 3.182$ [From table value]

So, 95% confidence interval of the predicted value of sales revenue corresponding to advertising expenditure of ₹ 600 would be :

$$8.20 \pm 3.182 \cdot \sqrt{2.93\left[1+\frac{1}{5}+\frac{(6-3)^2}{10}\right]}$$

or, 8.20 ± 7.891 or 0.309 and 16.091

$\Rightarrow$ ₹ 309 and ₹ 16091

Thus 95% confidence interval of predicted sales revenue ($\hat{Y}_0$) corresponding to advertising expenditure of ₹ 600 would be ₹ 309 and ₹ 16091.

(vi) 95% [= $100(1-\alpha)$% when $\alpha = 0.05$] confidence interval of expected sales revenue [$E(\hat{Y}_0)$] when advertising expenditures are ₹ 600 would be

$$\hat{Y}_0 \pm t_{\alpha/2, n-2}\sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}\left[\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum x_i^2}\right]}$$

[Here $\alpha = 0.05$, $t_{\alpha/2, n-2} = t_{0.025, 3} = 3.182$ [from table value]

$n = 5$, $X_0 = 6$, $\sum x_i^2 = 10$, $\bar{X} = 3$, $\sum_{i=1}^{n} e_i^2 \Big/ (n-2) = 2.93$

The prediction is still given by

$\hat{Y}_0 = 1.0 + 1.2X$, so that when $X = X_0 = 6$, $\hat{Y}_0 = 1.0 + 1.2 \times 6 = 8.20$]

i.e., $8.20 \pm 3.182\sqrt{2.93\left[\frac{1}{5}+\frac{(6-3)^2}{10}\right]}$

or, $8.20 \pm 3.182 \times 1.795$

or, 8.20 ± 5.72 or, 2.48 and 13.92

$\therefore$ 95% confidence interval for the average sales $[E(\hat{Y}_0)]$ corresponding to advertising expenditures ₹ 600 would be ₹ 2480 and ₹ 13920.

It should be noted that this confidence interval is narrower than the one we obtained for $\hat{Y}_0$.

Example 2.14. The following table (Table 2.6) gives data on the level of education (number of years of schooling), the mean hourly wages earned by the people at each level of education and the number of people at the stated level of education.

Table 2.6. Mean Hourly wage by Education

Years of schooling (X)	Mean hourly wage in \$ (Y)	Number of people
6	4.4567	3
7	5.7700	5
8	5.9787	15
9	7.3317	12
10	7.3182	17
11	6.5844	27
12	7.8182	218
13	7.8351	37
14	11.0223	56
15	10.6738	13
16	10.8361	70
17	13.6150	24
18	13.5310	31
Total		528

(i) Assuming a linear regression line of the form $Y_i = \alpha + \beta X_i + u_i$ $[u_i \sim N(0, \sigma_u^2)]$, find the OLS estimators of α and β.

(ii) Find var ($\hat{\alpha}$) and var ($\hat{\beta}$)

(iii) Find $SE(\hat{\alpha})$ and $SE(\hat{\beta})$

(iv) Find R^2

(v) Find $\sum_{i=1}^{n} e_i^2$

(vi) Predict/Forecast about the mean hourly wage when the level of education (years of schooling) is 20.

(vii) Construct 95% confidence interval for the point predictor of hourly wage $(\hat{Y}_0)$ when the level of education is, $X_0 = 20$.

(viii) Construct 95% confidence interval of expected mean hourly wage rate $[E(\hat{Y}_0)]$ when the level of education is, $X_0 = 20$.

Calculations for the Regression

Observations	Y_i	X_i	$x_i = X_i - \bar{X}$	$y_i = Y_i - \bar{Y}$	x_i^2	$x_i y_i$	$e_i = Y_i - \hat{Y}_i$
1	4.4567	6	−6	−4.218	36	25.308	0.291406
2	5.77	7	−5	−2.9047	25	14.5235	0.853137
3	5.9787	8	−4	−2.696	16	10.784	0.310268
4	7.3317	9	−3	−1.343	9	4.029	0.911699
5	7.3182	10	−2	−1.3565	4	2.713	0.14663
6	6.5844	11	−1	−2.0903	1	2.0903	−1.33874
7	7.8182	12	0	−0.8565	0	0	−0.85651
8	7.8351	13	1	−0.8396	1	−0.8396	−1.59118
9	11.0223	14	2	2.3476	4	4.6952	0.844454
10	10.6738	15	3	1.9991	9	5.9973	−0.25562
11	10.8361	16	4	2.1614	16	8.6456	−0.84483
12	13.615	17	5	4.9403	25	24.7015	1.182447
13	13.531	18	6	4.8563	36	29.1378	0.346878
Total	$\sum_{i=1}^{n} Y_i$ $= 112.7712$	$\sum_{i=1}^{n} X_i$ $= 156$	$\sum_{i=1}^{n} x_i$ $= 0$	$\sum_{i=1}^{n} y_i$ $= 0$	$\sum_{i=1}^{n} x_i^2$ $= 182$	$\sum_{i=1}^{n} x_i y_i$ $= 131.7856$	$\sum_{i=1}^{n} e_i$ $= 0$

Note : $n = 13$, $\therefore \bar{Y} = \sum_{i=1}^{n} Y_i \Big/ n = \dfrac{112.7712}{13} = 8.6747$

and $\bar{X} = \sum_{i-}^{n} X_i \Big/ n = \dfrac{156}{13} = 12$, $\sum_{i=1}^{n} X_i^2 = (6)^2 + (7)^2 + ... + (18)^2 = 2054$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \frac{131.7856}{182.0} = 0.7240967$$

and $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 8.6747 - 0.7240967 \times 12 = -0.01445$

$\therefore$ The estimated regression equation is

$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ or, $\hat{Y}_i = -0.01445 + 0.7240967\, X_i$

$\therefore e_i = Y_i - \hat{Y}_i = Y_i - 0.01445 - 0.7240967\, X_i$

Now different values of e_i can be obtained by taking different pairs of values of X_i and Y_i.

When $Y_i = 4.4567$ and $X_i = 6$, then from the above relation we get $e_i = 0.291406$. Similarly, other values of e_i are calculated.

Now, $\sum_{i=1}^{n} e_i^2 = (0.291406)^2 + (0.853137)^2 + ... + (0.346878)^2 = 9.83017$

and $\sum_{i=1}^{n} e_i^2 \Big/ (n-2) = \frac{9.83017}{13-2} = 0.8936$

(i) OLS estimates of α and β would be :

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \frac{131.7856}{182} = 0.7240967$$

and $\hat{\alpha} = \hat{Y} - \hat{\beta}\overline{X} = 8.6747 - 0.7240967 \times 12 = -0.01445$

$\therefore \hat{\alpha} = -0.0144$ and $\hat{\beta} = 0.7240$

(ii) We have to calculate the values of var ($\hat{\alpha}$) and var ($\hat{\beta}$)

We know that var ($\hat{\alpha}$) = $\frac{\sigma_u^2 \sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} x_i^2}$. Here σ_u^2 is unknown and hence it is replaced by its unbiased estimator $\hat{\sigma}_u^2 = \sum_{i=1}^{n} e_i^2 \Big/ (n-2) = 0.8936$

$$\therefore \text{var}(\hat{\alpha}) = \frac{\hat{\sigma}_u^2 \cdot \sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} x_i^2} = \frac{0.8936 \times 2054}{13 \times 182} = \frac{1835.4544}{2366} = 0.7757$$

$\therefore \text{var}(\hat{\alpha}) = 0.7757$

Again, $\text{var}(\hat{\beta}) = \frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2}$. Here σ_u^2 is unknown and hence it is replaced by its unbiased estimator $\hat{\sigma}_u^2 = \sum_{i=1}^{n} e_i^2 \Big/ (n-2) = 0.8936$

$$\therefore \text{var}(\hat{\beta}) = \frac{\hat{\sigma}_u^2}{\sum_{i=1}^{n} x_i^2} = \frac{0.8936}{182} = 0.004910 \quad \therefore \text{var}(\hat{\beta}) = 0.004910$$

(iii) We know that $SE(\hat{\alpha})=\sqrt{\text{var}(\hat{\alpha})}$ and $SE(\hat{\beta})=\sqrt{\text{var}(\hat{\beta})}$

$$\therefore SE(\hat{\alpha})=\sqrt{\text{var}(\hat{\alpha})}=\sqrt{0.7757}=0.8807$$

$$\text{and } SE(\hat{\beta})=\sqrt{\text{var}(\hat{\beta})}=\sqrt{0.004910}=0.07007$$

$$\therefore SE(\hat{\alpha})=0.8807 \text{ and } SE(\hat{\beta})=0.07007$$

(iv) Since $R^2=\dfrac{\text{ESS}}{\text{TSS}}=\dfrac{\hat{\beta}^2\cdot\sum_{i=1}^{n}x_i^2}{\sum_{i=1}^{n}y_i^2}$ $\therefore R^2=\dfrac{(0.7240)^2\times 182}{105.23017}=\dfrac{95.4000}{105.23017}=0.9065$

[Since $\sum_{i=1}^{n}y_i^2=\hat{\beta}^2\cdot\sum_{i=1}^{n}x_i^2+\sum_{i=1}^{n}e_i^2$ or, TSS = ESS + RSS

$$\therefore \sum_{i=1}^{n}y_i^2=(0.7240)^2\times 182+9.83017\left[\because \sum_{i=1}^{n}e_i^2=9.83017\right]$$

$$=95.4000+9.83017=105.23017]$$

(v) Now, $\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2=(0.291406)^2+(0.853137)^2+\ldots+(0.346878)^2$

$$=9.83017 \quad \therefore \sum_{i=1}^{n}e_i^2=9.83017.$$

(v) We have to forecast/predict about the mean hourly wage rate when the level of education (years of schooling) is 20.

Since the estimated regression equation is $\hat{Y}_i=-0.0144+0.7240\,X_i$

The point predictor of Y is given by $\hat{Y}=\hat{\alpha}+\hat{\beta}X$.

When $X=X_0$, $\hat{Y}_0=\hat{\alpha}+\hat{\beta}X_0$.

So, when the level of education is $X_0=20$, the mean hourly wage rate would be $\hat{Y}_0=-0.0144+0.7240\times 20=14.4656$

So, mean hourly wage rate would be $ 14.4656 when years of schooling increases to 20.

(vi) We have to construct 95% confidence interval for the point predictor of hourly wage rate (Y_0) when the level of education becomes $X_0=20$. This confidence interval would be

$$\hat{Y}_0\pm t_{0.025,n-2}\sqrt{\frac{\sum_{i=1}^{n}e_i^2}{n-2}\left[1+\frac{1}{n}+\frac{(\overline{X}-X_0)^2}{\sum_{i=1}^{n}x_i^2}\right]}$$

When $X = X_0 = 20$, $Y_0 = \hat{Y}_0 = -0.0144 + 0.7240 \times 20 = 14.4656$

$t_{\alpha/2, n-2} = t_{0.025,13-2} = t_{0.025,11} = 2.201$ (Table value) as $\alpha = 0.05$, $n = 13$.

Again, $X = 20$, $\bar{X} = 8.6747$, $\sum_{i=1}^{n} x_i^2 = 182$, $\sum_{i=1}^{n} e_i^2 \Big/ (n-2) = 0.8936$

$\therefore$ $100(1-\alpha) = 95\%$ (as $\alpha = 0.05$) confidence interval of Y_0 would become,

$$\hat{Y}_0 \pm t_{0.025,11} \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}\left[1 + \frac{1}{n} + \frac{(\bar{X} - X_0)^2}{\sum x_i^2}\right]}$$

$$\text{or, } 14.4656 \pm 2.201 \sqrt{0.8936\left[1 + \frac{1}{13} + \frac{(12-20)^2}{182}\right]}$$

$$\text{or, } 14.4656 \pm 2.201 \sqrt{0.8936\left[1 + \frac{1}{13} + \frac{64}{182}\right]}$$

$$\text{or, } 14.4656 \pm 2.201 \sqrt{0.8936[1 + 0.07692 + 0.35164]}$$

$$\text{or, } 14.4656 \pm 2.201 \sqrt{0.8936 \times 1.42856} \text{ or, } 14.4656 \pm 2.201 \sqrt{1.27656}$$

or, $14.4656 \pm 2.201 \times 1.12984$ or, 14.4656 ± 2.4868

i.e. 11.9788 and 16.9524

$\therefore$ 95% confidence interval of hourly wage rate would be \$ 11.9788 and \$ 16.9524 when the level of education (years of schooling) is 20.

(vii) We have to construct 95% confidence interval of expected mean hourly wage rate when the level of education is $X_0 = 20$ (years of schooling).

When $X = X_0 = 20$, $E(Y/X_0 = 20)$ can be obtained as $\hat{Y}_0 = \hat{\alpha} + \hat{\beta} X_0 = -0.0144 + 0.7240 \times 20 = 14.4656$.

Thus 95% confidence interval of $E(Y/X_0)$ when $X_0 = 20$ would be :

$$\hat{Y}_0 \pm t_{0.025,n-2} \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}\left(\frac{1}{n} + \frac{(\bar{X} - X_0)^2}{\sum_{i=1}^{n} x_i^2}\right)}$$

Here $\hat{Y}_0 = 14.4656$, $n = 13$, $t_{0.025,\, n-2} = t_{0.025,11} = 2.201$ (Table value)

$\bar{X} = 12$, $X_0 = 20$ and $\sum_{i=1}^{n} x_i^2 = 182$, $\sum_{i=1}^{n} e_i^2 \Big/ (n-2) = 0.8936$

$$\text{Now, } \hat{Y}_0 \pm t_{0.025,n-2}\sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}\left[\frac{1}{n}+\frac{(\bar{X}-X_0)^2}{\sum_{i=1}^{n} x_i^2}\right]}$$

$$= 14.4656 \pm 2.201\sqrt{0.8936\left[\frac{1}{13}+\frac{(12-20)^2}{182}\right]}$$

or, $14.4656 \pm 2.201\sqrt{0.8936 \times 0.42857}$ or, $14.4656 \pm 2.201\sqrt{0.382970}$

or, $14.4656 \pm 2.201 \times 0.61878$ or, 14.4656 ± 1.3620

i.e., 13.1036 and 15.8276

Thus, 95% confidence interval of expectred mean hourly wage rate corresponding to the level of education (year of schooling) $X_0 = 20$ would be \$13.1036 and \$15.8276.

Example 2.15. The following table (Table 2.7) shows consumption expenditure and income (in billions of \$) of a country over the period 2007-2018.

Table 2.7. Consumption expenditure and income of a country (in billions of \$)

Year	Consumption expenditure (Y)	Income (X)
2007	282.3	359.9
2008	291.1	370.1
2009	312.3	394.7
2010	326.5	414.5
2011	336.6	430.8
2012	356.6	458.7
2013	376.6	483.5
2014	402.9	415.5
2015	434.7	557.4
2016	468.3	613.1
2017	494.3	658.9
2018	538.9	721.0

From the data given in the table we have the following results :

$$\hat{Y}_t = 31.76 + 0.71\,X_t,\quad R^2 = 0.998,\quad \hat{\sigma}_u^2 = \sum_{i=1}^{n} e_i^2 \Big/ (n-2) = 285.61$$

$$(5.39)\quad (0.01)$$

$\bar{X} = 498$, $n = 12$

$$\sum_{i=1}^{n}(X_t-\bar{X})^2 = \sum_{i=1}^{n} x_i^2 = 151{,}482,\ (\bar{X}-X_0)^2 = 123{,}904$$

X_0= Income in the year 2025 = \$850 billion.

(i) Forecast about consumption expenditure of the country for the year 2025 if income in that year increases to \$850 billion.

(ii) Construct 95% confidence interval of the consumption expenditure (Predicted/forecasted) for the year 2025.

Solution :

(i) Since $\hat{Y}_i = 31.76 + 0.71\, X_i$, so, when $X_i = 850 = X_0$

$$\hat{Y}_{2025} = 31.76 + 0.71 \times 850 = 635.26 \approx \$635 \text{ billion.}$$

So, consumption expenditure in 2025 will be $635 billion when the income of the country becomes $850 billion.

(ii) 95% confidence interval of the predicted consumption expenditure of the country for the year 2025 would be :

$$\hat{Y}_{2025} \pm t_{0.025,n-2} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}\left(1+\frac{1}{n}+\frac{(\bar{X}-X_0)^2}{\sum_{i=1}^{n} x_i^2}\right)}$$

or, $$635 \pm t_{0.025,10}\sqrt{285.61\left[1+\frac{1}{12}+\frac{123{,}904}{151{,}482}\right]}$$

or, $$635 \pm 2.228\sqrt{285.61 \times 1.9014}$$

or, $$635 \pm 2.228\sqrt{534.05885}$$

or, $$635 \pm 2.228 \times 23.1097$$

or, $$635 \pm 51.4884$$

i.e., 583.5116 and 686.4884

Thus 95% confidence interval of predicted consumption expenditure of the country for the year 2025 would be $583.5116 billion and $686.4884 billion.

EXERCISE

1. In a simple linear regression model $Y_i = \alpha + \beta X_i + u_i$ for $i = 1, 2, ..., n$ why do we insert the random disturbance term 'u' ?
2. State and explain the assumptions of a classical linear regression model (CLRM).
3. In a simple linear regression model of the form $Y_i = \alpha + \beta X_i + u_i$, $i = 1, 2, ..., n$ how can you estimate the regression parameters α and β ?
4. Describe briefly the method of moments, used in estimating the regression parameters in a two variable linear regression model.
5. Describe briefly the method of least squares used in estimating the regression parameters relating to a two variable linear regression model.
6. How can you estimate a linear function (two variable) whose intercept is zero ?
7. How can you estimate the elasticities from an estimated regression line ?
8. State and prove the properties of the least squares estimators relating to a two variable linear regression model (CLRM).
9. Show that in a classical linear regression model the estimated regression parameters are unbiased.
10. Determine the mean and variance of $\hat{\alpha}$ and $\hat{\beta}$ relating to a model $Y_i = \alpha + \beta X_i + u_i$ for $i = 1, 2, ..., n$.

Estimate α and β and calculate estimates of variances of your estimates. Estimate the conditional mean value of Y corresponding to a value of X fixed at $X = 20$.

27. The following table shows the investment expenditure and the long run interest rate over the ten year period :

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Investment	656	804	836	765	777	711	755	747	696	787
Interest	0.05	0.045	0.045	0.055	0.06	0.06	0.06	0.05	0.07	0.065

Test the hypothesis that investment is interest elastic by fitting a regression line to the above data and conducting the relevant tests of significance.

28. Given the following data :

$\Sigma x_i y_i = 200,\ \ \Sigma x_i^2 = 100,\ \ \Sigma y_i^2 = 500,\ \ \bar{X} = 100,\ \ \bar{Y} = 150,\ n = 27$, estimate the parameters in the model : $Y_i = \alpha + \beta X_i + u_i$ and test the hypothesis $H_0 : \beta = 1.5$ against the alternative, (i) $\beta \neq 1.5$, (ii) $\beta > 1.5$, (iii) $\beta < 1.5$.

29. The true relationship between X and Y in the population is given by :

$Y_i = 2 + 3X_i + u_i$. Suppose, the value of X in the sample of 10 observations are 1, 2, 3, ..., 10. The values of the disturbances are drawn at random from a normal population with zero mean and constant variance :

$u_1 = 0.464$, $u_2 = 0.06$, $u_3 = 1.48$, $u_4 = 1.02$, $u_5 = 1.39$, $u_6 = 0.91$, $u_7 = -1.50$, $u_8 = -0.69$, $u_9 = 0.18$ and $u_{10} = -1.37$.

(i) Present the 10 observed values of X and Y.

(ii) Estimate the least squares estimates of the regression coefficients and their standard errors.

(iii) Obtain the predicted value of Y for $X = 12$.

30. The following data gives the production of coal and the number of wage earners in the coal industry :

Output : (million tonnes)	210.8	210.1	211.5	208.9	207.4	205.3	198.8	192.1	183.2	176.8
Number of workers : (000's)	706.2	703.1	701.8	699.1	697.4	795.3	692.7	630.2	602.1	531.0

(i) Estimate the production function (linear) of coal.

(ii) Find average and marginal productivity of labour.

(iii) Estimate t-ratios and test their significance.

31. The following are data on

Y = Quit rate per 100 employees in manufacturing

X = unemployment rate

The data are for the United States and cover the period 1960-1972.

Year	Y	X	Year	Y	X
1960	1.3	6.2	1966	2.6	3.2
1961	1.2	7.8	1967	2.3	3.6
1962	1.4	5.8	1968	2.5	3.3
1963	1.4	5.7	1969	2.7	3.3
1964	1.5	5.0	1970	2.1	5.6
1965	1.9	4.0	1971	1.8	6.8
			1972	2.2	5.6

11. State and prove the Gauss-Markov theorem, with reference to a CLRM.

12. What is meant by the term BLUE ? Show that (i) $\hat{\alpha}$ is the BLUE of α and (ii) $\hat{\beta}$ is the BLUE of β in the CLRM, $Y = \alpha + \beta X + u$.

13. How can you determine the variance of the random distrubance term in the model $Y = \alpha + \beta X + u$.

14. Show that $\sum_{i=1}^{n} e_i^2 \Big/ (n-2)$ is an unbiased estimator of the variance of the random disturbance term i.e., σ_u^2 in the model $Y_i = \alpha + \beta X_i + u_i$.

15. What is a maximum likelihood estimator (MLE) ? Show that in the model $Y_i = \alpha + \beta X_i + u_i$ (i) $\hat{\alpha}$ is the MLE of α.

    (ii) $\hat{\beta}$ is the MLE of β.

    (iii) $\sum_{i=1}^{n} e_i^2 \Big/ n$ is the MLE of σ_u^2.

16. Show that the least squares estimates $\hat{\alpha}$ and $\hat{\beta}$ are such that :

$$\text{(i) } \hat{\alpha} - N\left[\alpha, \sigma_u^2\left(\frac{1}{n} + \frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2}\right)\right] \qquad \text{(ii) } \hat{\beta} \sim N\left[\beta, \frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2}\right]$$

17. Describe the testing procedure of the significance of the regression coefficients of the model $Y_i = \alpha + \beta X_i + u_i$ for $i = 1, 2, \ldots, n$.

18. What is meant by goodness of fit of the correlation coefficient R^2 ?

19. Show that Total sum of squares = Explained sum of squares + Unexplained sum of squares.

20. What is coefficient of determination ? Show that it lies between 0 and 1 and hence show that the value of correlation coefficient between two variables lies between –1 and +1.

21. How can you formally write the regression results of the regression model $Y_i = \alpha + \beta X_i + u_i$ where u_i (for $i = 1, 2, \ldots, n$) satisfies all the properties of CLRM ?

22. How can you use the analysis of variance in the simple classical linear regression model?

23. What is the meaning of the term 'Prediction' ? How can you incorporate the term in the CLRM ? Distinguish between point prediction and interval prediction in this regard.

24. Show that the *OLS* point predictor in the CLRM satisfies the BLUE property.

25. The following sums where obtained from 16 pairs of observations on X and Y : $\Sigma Y_i^2 = 526$, $\Sigma X_i^2 = 657$, $\Sigma X_i Y_i = 492$, $\Sigma Y_i = 63$, $\Sigma X_i = 96$, Estimate the parameters in the model : $Y_i = \alpha + \beta X_i + u_i$ and R^2.

    Test the hypothesis that $\beta = 2.0$.

26. A sample of 20 observations corresponding to the regression model : $Y_i = \alpha + \beta X_i + u_i$ gave the following data :

$$\Sigma Y_i = 21.9,\ \Sigma(Y_i - \bar{Y})^2 = 86.9,\ \Sigma(X_i - \bar{X})(Y_i - \bar{Y}) = 106.4,\ \Sigma X_i = 186.2,\ \Sigma(X_i - \bar{X})^2 = 215.4.$$

(i) Calculate a regression of Y on X, $Y = \alpha + \beta X + u$

(ii) Construct a 95% confidence interval for β.

(iii) Test the hypothesis $H_0 : \beta = 0$ against the alternative $H_1 : \beta \neq 0$ at the 5% level of significance.

32. Using the time series in the following table, estimate the consumption function and the saving function (in linear form) of the UK. What is the marginal propensity to consume and the marginal propensity to save of the country ? Interpret the intercepts of the two functions.

Year	Income (in £m)	Consumption (in £m)
1964	26,934	21,439
1965	28,729	22,833
1966	30,171	24,205
1967	31,781	25,307
1868	33,450	27,020

33. A random sample of ten families had the following income and food expenditure (in £ per week) :

Families	:	A	B	C	D	E	F	G	H	I	J
Family Income	:	200	300	330	400	150	130	260	380	350	430
Family Expenditure	:	70	90	80	110	50	40	80	100	90	100

Estimate the regression line of food expenditure on income and interpret your results.

34. The following results have been obtained from a sample of 11 observations on the value of sales (Y) of a firm and the corresponding prices (X).

$$\bar{X} = 519.18,\ \bar{Y} = 217.82,\ \Sigma X_i^2 = 3{,}134{,}543,\ \Sigma X_i Y_i = 1{,}296{,}836,\ \Sigma Y_i^2 = 539{,}512$$

(i) Estimate the regression line of sales on price and interpret the results.

(ii) What is the part of the variation in sales which is not explained by the regression line ?

(iii) Estimate the price elasticity of sales.

35. The following table gives the quantities of commodity Z bought in each year from 2009-2018 and the corresponding prices.

Year	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Quantity (in tons)	770	785	790	795	800	805	810	820	840	850
Price (in £)	18	16	15	15	12	10	10	7	9	6

(i) Estimate the linear demand function for commodity Z.

(ii) Calculate the price elasticity of demand.

(iii) Forecast the demand at the mean price of the sample.

(iv) Forecast the demand at P = 20.

36. A sample of 20 observations in a time series data on X and Y is to be used for estimating the linear function $Y = \alpha + \beta X + u$. The first 10 observations yield the following results;

$$\bar{X} = 15.30,\ \bar{Y} = 160.00,\ \sum_{i=1}^{10}(X_i - \bar{X})^2 = 78.00,\ \sum_{i=1}^{10}(Y_i - \bar{Y})^2 = 45{,}600,$$

$$\sum_{i=1}^{10}(X_i - \bar{X})(Y_i - \bar{Y}) = -1568.00$$

The 10 subsequent pairs of values of X and Y yield : $\bar{X} = 14.08$, $\bar{Y} = 106.00$,

$$\sum_{i=11}^{20}(X_i - \bar{X})^2 = 98.16,\quad \sum_{i=11}^{20}(Y_i - \bar{Y})^2 = 62,440,\quad \sum_{i=11}^{20}(X_i - \bar{X})(Y_i - \bar{Y}) = -2308.80.$$

Has the function changed over the two decades ?

37. The following table includes data on the quantity supplied for exports of commodity x and its price.

Year	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Quantity (million tons) (Y) :	50	40	30	40	70	90	80	100	80	20
Export price (X) : ($ per ton)	20	40	20	30	80	70	60	80	70	30

(i) Test the hypothesis that the quantity supplied and price are positively related, by estimating the export supply function $Y = \alpha + \beta X + u$. Interpret your results.

(ii) Show that $\hat{\beta}$ is a part of the price elasticity of supply, and obtain a numerical value for the latter.

(iii) If price in 2024 becomes $80 and in 2034 it becomes $120 then forecast the export volume of the commodity in these years.

38. The following table includes the total cost and the level of output of firm A over a ten-year period.

Year	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Quantity (X) : (0000 units)	40	42	48	55	65	79	88	100	120	140
Total cost (Y) : (0000 dollars)	150	140	160	170	150	162	185	165	190	185

(i) Estimate the linear cost function $Y = \alpha + \beta X$.

(ii) Find the AVC, MC and AC, and plot them roughly on a graph.

39. The total investment function for the economy as a whole is assumed to be of the form:

$I = \alpha r^{\beta}.e^{u}$

where I = investment, r = rate of interest.

The following sample is given :

I ($ billion) :	9.0	5.5	8.5	4.0	3.5	2.5	3.0	1.5	1.2	1.8	1.5
r (percent) :	2	3	2	4	5	6	4	6	8	7	9

(i) Estimate the parameters of the investment function by OLS.

(ii) Test the statistical significane of the coefficients at 1% level of significance.

(iii) Construct a 95% confidence interval for β.

(iv) Find the value of R^2 and interpret the result.

40. (a) Write down the assumptions essential for each of the following tasks :

(i) Proving that the OLS estimators are unbiased.

(ii) Proving that the OLS estimators are efficient.

(iii) Carrying out the t and F tests.

(b) Given the data on Y and X explain how you will estimate the parameters in the following models by using OLS technique :

(i) $Y = \alpha X^{\beta}$, (ii) $Y = \dfrac{X}{(\alpha X - \beta)}$, (iii) $Y = \alpha + \beta \log X$

41. Consider the following estimated regression equation : $\hat{Y}_i = \hat{\alpha} + 1.5X_i + e_i$ with estimated standard error of β coefficient 0.5. It has further been given that $R^2 = 0.5$, $\bar{X} = 10$,

$\bar{Y} = 25$ and $\sum_{i=1}^{n} Y_i^2 = 6895$

Find out the following :

(a) Sample size (n), (b) The estimated intercept ($\hat{\alpha}$), (c) Total sum of squares (TSS),

(d) Residual sum of squares (RSS), (e) Estimated error variance ($\hat{\sigma}_u^2$).

42. (a) In a two variable regression model, show that $R^2 = \dfrac{\text{ESS}}{\text{TSS}}$.

(b) Explain why random error term is introduced in an econometric model.

(c) Which one would you consider to be CLRM ? Give reasons.

(i) $Y_i = \beta_1 + \beta_2 X_i^2$, (ii) $Y_i = \beta_1 + \sqrt{\beta_2 X_i}$, (iii) $Y_i = \beta_1 + \beta_2 X_i^{\beta_3}$, (iv) $Y_i = \beta_1 + \beta_2 X_i$

43. Consider the following estimated two variable CLRM : $Y_i = \alpha + 0.5X_i + u_i$ when $n = 22$,

$\bar{X} = 10$, $\bar{Y} = 15$, $\Sigma X_i^2 = 2201$, $\Sigma Y_i^2 = 4951$.

(i) Obtain the estimated regression coefficient when X is regressed on Y.

(ii) Obtain the coefficient of correlation between X and Y.

(iii) Obtain the unbiased estimate of the error variance when Y is regressed on X.

(iv) Obtain the estimated value of the intercept term and its estimated standard error when Y is regressed on X.

(v) Test the suggestion that Y is positively related to X at 5% level of significance.

44. Consider the following regression equation $Y_i = \alpha + \beta X_i + u_i$, where $n = 10$, $\Sigma X_i = 70$,

$\Sigma Y_i = 80$, $\Sigma X_i^2 = 600$, $\Sigma Y_i^2 = 734$, $\Sigma X_i Y_i = 480$.

(a) Obtain the estimated value of α and β.

(b) Test the hypothesis that X and Y are negatively correlated against the hypothesis that they are not at 5% level of significance.

3

Multiple Linear Regression Model

3.1. Introduction

In simple regression analysis we study the relationship between an explained (dependent) variable Y and an explanatory (independent) variable X. In multiple regression analysis we study the relationship between Y and a number of explanatory variables $X_1, X_2, ..., X_K$. For example, in demand studies we may be interested in investigating the relationship between quantity demanded of a good and price of that good, prices of substitute goods and income of the consumer.In fact this problem can be analysed with the help of multiple regression analysis.

Let us consider a linear regression model where there are K independent variables say $X_1, X_2, ..., X_K$ and Y is the only dependent variable. In this case the regression model is given by,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_K X_{Ki} + u_i \text{ where } i = 1, 2, ..., n \qquad ... (3.1)$$

Here $\beta_0, \beta_1, \beta_2, ..., \beta_K$ are $(K + 1)$ regression parameters.

[K = number of explanatory variables and $K + 1$ = number of regression parameters]

u = random disturbance term / error term.

β_0 = constant term and $\beta_1, \beta_2, ..., \beta_K$ are the partial regression coefficients.

We make the following assumptions about u_i :

(i) $E(u_i) = 0$ for all i, $i = 1, 2, ..., n$

(ii) $\text{var}(u_i) = \sigma_u^2$ for all i

(iii) u_i and u_j are independent for all $i \neq j$

(iv) u_i and X_j are independent for all i and j

(v) u_i is normally distributed for all i [$u_i \sim N(0, \sigma_u^2)$]

(vi) There is no linear dependencies in the explanatory variables.

Under the first four assumptions, we can show that the method of least squares gives estimators of $\beta_0, \beta_1, \beta_2, ..., \beta_K$ that are unbiased and have minimum variance.

In equation (3.1) if we put $i = 1, 2, ..., n$, we have :

For $i = 1$, $Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + ... + \beta_K X_{K1} + u_1$

" $i = 2$, $Y_2 = \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + ... + \beta_K X_{K2} + u_2$

" $i = 3$, $Y_3 = \beta_0 + \beta_1 X_{13} + \beta_2 X_{23} + ... + \beta_K X_{K3} + u_3$

....

" $i = n$, $Y_n = \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + ... + \beta_K X_{Kn} + u_n$

In vector-matrix form the above set of n equations can be written as :

$$Y = X\beta + u \qquad \text{.... (3.2)}$$

where

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1}, X = \begin{bmatrix} 1 & X_{11}X_{21}\cdots X_{K1} \\ 1 & X_{12}X_{22}\cdots X_{K2} \\ \vdots & \vdots \\ 1 & X_{1n}X_{2n}\cdots x_{Kn} \end{bmatrix}_{n\times(K+1)}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ \beta_K \end{bmatrix}_{(K+1)\times 1} \text{ and } u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}_{n\times 1}$$

Equation (3.2) represents a set of n equations where there are K independent (explanatory) variables with n units of sample size.

The model is said to be a Classical Linear Regression Model (CLRM) if it satisfies the following properties :

(i) $E(u)$ is a null vector

$$\text{where } E(u) = E\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} E(u_1) \\ E(u_2) \\ \vdots \\ E(u_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{n\times 1} = O_{n\times 1}$$

since $E(u_i) = 0$ for all i, $i = 1, 2, ..., n$

(ii) The dispersion matrix (variance covariance matrix) of the disturbance vector u is a scalar matrix.

$$D(u) = \sigma_u^2 . I_n \text{ where } I_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}_{n\times n} = \text{Identity matrix and } \sigma_u^2 \text{ is a constant.}$$

Proof : Let us suppose u is a random vector variable, then dispersion matrix, $D(u)$ $= E[u - E(u)][u - E(u)]'$

$= E[uu']$, as $E(u) = 0$

$$= E\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_n \end{bmatrix}[u_1, u_2, u_3, ..., u_n]$$

$$= E\begin{bmatrix} u_1^2 & u_1u_2 & \cdots & \cdots & u_1u_n \\ u_2u_1 & u_2^2 & \cdots & \cdots & u_2u_n \\ u_3u_1 & u_3u_2 & u_3^2 & \cdots & u_3u_n \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ u_nu_1 & u_nu_2 & \cdots & \cdots & u_n^2 \end{bmatrix} = \begin{bmatrix} E(u_1^2) & E(u_1u_2) & \cdots & \cdots & E(u_1u_n) \\ E(u_2u_1) & E(u_2^2) & \cdots & \cdots & E(u_2u_n) \\ E(u_3u_1) & E(u_3u_2) & E(u_3^2) & \cdots & E(u_3u_n) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ E(u_nu_1) & E(u_nu_2) & \cdots & \cdots & E(u_n^2) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_u^2 & 0 & \cdots & \cdots & 0 \\ 0 & \sigma_u^2 & \cdots & \cdots & 0 \\ 0 & 0 & \sigma_u^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \sigma_u^2 \end{bmatrix} \qquad \begin{array}{l} \text{since } E(u_i u_j) = 0 \text{ for } i, j, i \neq j \\ \text{and } E(u_i^2) = \sigma_u^2 \text{ for all } i, i = 1, 2, \ldots, n \end{array}$$

$$\text{Now, } D(u) = \sigma_u^2 \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & \cdots & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} = \sigma_u^2 I_n$$

$\therefore D(u) = \sigma_u^2 I_n$

(iii) u is a multivariate normal vector with mean $O_{n\times n}$ and variance-covariance matrix $D(u)$.

(iv) All the independent variables $X_1, X_2, \ldots, X_K$ are non-stochastic or nonrandom or X is a non-stochastic matrix.

(v) Rank of matrix X is $(K + 1)$

Rank of a matrix implies the maximum number of linearly independent columns of the matrix.

Since X is a matrix of order $n \times (K + 1)$, all the columns of X should be linearly independent. Now X' is a matrix of order $(K + 1) \times n$ and $(X'X)$ is a matrix of order $(K + 1)(K + 1)$. If the rank of X is $(K + 1)$ then rank of $(X'X)$ is also $(K + 1)$ and $|X'X| \neq 0$ if its rank is $(K + 1)$ and $|X'X| = 0$ if its rank is $<(K + 1)$

If $|X'X| = 0$, then $(X'X)^{-1}$ does not exist.

3.2. The Least Squares Method (OLS) for Estimation of Regression Parameters

In vector matrix form the general linear regression model (Equation 3.1) can be written as, $Y = X\beta + u$

$$\text{where } Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1}, X = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{K1} \\ 1 & X_{12} & X_{22} & \cdots & X_{K2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{Kn} \end{bmatrix}_{n\times(K+1)}$$

$$u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}_{n\times 1} \text{ and } \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}_{(K+1)\times 1}$$

Let $\hat{Y} = X\hat{\beta}$ be the vector of the regressed value of Y.

and $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \beta_k \end{bmatrix}_{(K+1)\times 1}$ be the vector of estimators,

where $\hat{y}$ is a $(n \times 1)$ order vector and X is a $n \times (K+1)$ order matrix.

Let e be the residual vector, i.e., $e = Y - \hat{Y}$ where $Y = X\beta + u$ and $\hat{Y} = X\hat{\beta}$

$\therefore e = Y - X\hat{\beta}$

Here, $e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}_{n\times 1}$ and $e' = [e_1, e_2, \ldots, e_n]_{1\times n}$

Now $e'e = [e_1, e_2, \ldots, e_n] \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = e_1^2 + e_2^2 + \ldots + e_n^2 = \sum_{i=1}^{n} e_i^2 \quad \therefore e'e = \sum_{i=1}^{n} e_i^2$.

$$\therefore e'e = (Y - X\hat{\beta})'(Y - X\hat{\beta})$$
$$= Y'Y - \hat{\beta}'X'Y - Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}$$

Here, $\hat{\beta}'X'Y$ is scalar (1×1), it is equal to its transpose ; i.e. $\hat{\beta}'X'Y = Y'X\hat{\beta}$

Now by OLS method we have to minimise $\sum_{i=1}^{n} e_i^2 = e'e$ with respect to $\hat{\beta}$.

Now $\frac{d(e'e)}{d\hat{\beta}} = O - X'Y - X'Y + 2X'X\hat{\beta} = 0$

or, $2X'X\hat{\beta} = 2X'Y$ or, $(X'X)\hat{\beta} = X'Y$

or, $\hat{\beta} = (X'X)^{-1}X'Y$, where $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_K \end{bmatrix}_{(K+1)\times 1}$

To derive this result more clearly we consider a three variable (with two explanatory variables i.e., when $K = 2$) linear regression model which takes the form :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \quad i = 1, 2, \ldots, n$$

Now, $\bar{Y} = \beta_0 + \beta_1\bar{X}_1 + \beta_2\bar{X}_2 + \bar{u}$ where $\bar{u} = 0$

If we take deviations from respective means, we have

$$Y_i - \bar{Y} = \beta_1(X_{1i} - \bar{X}_1) + \beta_2(X_{2i} - \bar{X}_2) + (u_i - \bar{u})$$

or, $y_i = \beta_1 x_{1i} + \beta_2 x_2 i + u_i$, $i = 1, 2, \ldots, n$ where $x_{1i} = X_{1i} - \bar{X}_1$, $x_{2i} = X_{2i} - \bar{X}_2$, and $(u_i - \bar{u}) = u_i$ as $\bar{u} = 0$

Let us define $X = \begin{bmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ \vdots & \\ x_{1n} & x_{2n} \end{bmatrix}_{n\times 2}$

$$X' = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \end{bmatrix}_{2\times n}, \; Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1}, \; u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}_{n\times 1} \text{ and } \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}_{2\times 1}.$$

In vector-matrix form the equation,

$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$ for $i = 1, 2, \ldots, n$, can be written as $Y = X\beta + u$

$$\text{Now } X'X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \end{bmatrix} \begin{bmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ \vdots & \vdots \\ x_{1n} & x_{2n} \end{bmatrix} = \begin{bmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i}x_{2i} \\ \Sigma x_{1i}x_{2i} & \Sigma x_{2i}^2 \end{bmatrix}$$

$$\text{Now } \hat{\beta}'X'X\hat{\beta} = [\hat{\beta}_1, \hat{\beta}_2] \begin{bmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i}x_{2i} \\ \Sigma x_{1i}x_{2i} & \Sigma x_{2i}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$$

$$= [\hat{\beta}_1, \hat{\beta}_2] \begin{bmatrix} \hat{\beta}_1\Sigma x_{1i}^2 & + & \hat{\beta}_2\Sigma x_{1i}x_{2i} \\ \hat{\beta}_1\Sigma x_{1i}x_{2i} & + & \hat{\beta}_2\Sigma x_{2i}^2 \end{bmatrix}$$

$$= \hat{\beta}_1^2\Sigma x_{1i}^2 + \hat{\beta}_1\hat{\beta}_2\Sigma x_{1i}x_{2i} + \hat{\beta}_1\hat{\beta}_2\Sigma x_{1i}x_{2i} + \hat{\beta}_2^2\Sigma x_{2i}^2$$

$$\therefore \hat{\beta}'X'X\hat{\beta} = \hat{\beta}_1^2\Sigma x_{1i}^2 + 2\hat{\beta}_1\hat{\beta}_2\Sigma x_{1i}x_{2i} + \hat{\beta}_2^2\Sigma x_{2i}^2$$

$$\text{Now } \frac{d}{d\hat{\beta}}\left(\hat{\beta}'X'X\hat{\beta}\right) = \begin{bmatrix} \dfrac{d}{d\hat{\beta}_1}(\hat{\beta}'X'X\hat{\beta}) \\ \dfrac{d}{d\hat{\beta}_2}(\hat{\beta}'X'X\hat{\beta}) \end{bmatrix}$$

$$= \begin{bmatrix} 2\hat{\beta}_1\Sigma x_{1i}^2 + 2\hat{\beta}_2\Sigma x_{1i}x_{2i} \\ 2\hat{\beta}_1\Sigma x_{1i}x_{2i} + 2\hat{\beta}_2\Sigma x_{2i}^2 \end{bmatrix} = 2\begin{bmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i}x_{2i} \\ \Sigma x_{1i}x_{2i} & \Sigma x_{2i}^2 \end{bmatrix}\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = 2(X'X)\hat{\beta}$$

$$\therefore \frac{d}{d\hat{\beta}}(\hat{\beta}'X'X\hat{\beta}) = 2X'X\hat{\beta}$$

Again $Y'X\hat{\beta} = [y_1, y_2, \ldots, y_n] \begin{bmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ \vdots & \vdots \\ x_{1n} & x_{2n} \end{bmatrix}\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$

$$= [\Sigma y_i x_{1i} \;\; \Sigma y_i x_{2i}]\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \hat{\beta}_1\Sigma x_{1i}y_i + \hat{\beta}_2\Sigma x_{2i}y_i$$

Now $\frac{d}{d\hat{\beta}}(Y'X\hat{\beta}) = \begin{bmatrix} \frac{d}{d\hat{\beta}_1}(Y'X\hat{\beta}) \\ \frac{d}{d\hat{\beta}_2}(Y'X\hat{\beta}) \end{bmatrix} = \begin{bmatrix} \Sigma x_{1i}y_i \\ \Sigma x_{2i}y_i \end{bmatrix} = X'Y$

[since $[\Sigma x_{1i}y_i \;\; \Sigma x_{2i}y_i]' = (Y'X)' = X'Y$]

$$\therefore \frac{d}{d\hat{\beta}}(Y'X\hat{\beta}) = X'Y$$

Again $\hat{\beta}'X'Y = [\hat{\beta}_1 \; \hat{\beta}_2]\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$

$$= [\hat{\beta}_1 \; \hat{\beta}_2]\begin{bmatrix} \Sigma x_{1i}y_i \\ \Sigma x_{2i}y_i \end{bmatrix} = \hat{\beta}_1\Sigma x_{1i}y_i + \hat{\beta}_2\Sigma x_{2i}y_i$$

Now $\frac{d}{d\hat{\beta}}(\hat{\beta}'X'Y) = \begin{bmatrix} \frac{d}{d\hat{\beta}_1}(\hat{\beta}'X'Y) \\ \frac{d}{d\hat{\beta}_2}(\hat{\beta}'X'Y) \end{bmatrix} = \begin{bmatrix} \Sigma x_{1i}y_i \\ \Sigma x_{2i}y_i \end{bmatrix} = X'Y$

Thus we have $\frac{d}{d\hat{\beta}}(\hat{\beta}'X'X\hat{\beta}) = 2X'X\hat{\beta}$,

$\frac{d}{d\hat{\beta}}(Y'X\hat{\beta}) = X'Y$ and $\frac{d}{d\hat{\beta}}(\hat{\beta}'X') = X'Y$

Since $e'e = Y'Y - \hat{\beta}'X'Y - Y'X\hat{\beta} + \hat{\beta}X'X\hat{\beta}$

$$\therefore \frac{d}{d\hat{\beta}}(e'e) = 0 - X'Y - X'Y + 2X'X\hat{\beta} = 0$$

or, $2X'X\hat{\beta} = 2X'Y \quad \therefore \hat{\beta} = (X'X)^{-1}X'Y$.

Thus the result is derived more clearly in terms of a three variable linear regression model.

Thus we have, $\hat{\beta} = (X'X)^{-1}X'Y$

$$\text{or, } \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i}x_{2i} \\ \Sigma x_{1i}x_{2i} & \Sigma x_{2i}^2 \end{bmatrix}^{-1} \begin{bmatrix} \Sigma x_{1i}y_i \\ \Sigma x_{2i}y_i \end{bmatrix}$$

For $K = 2$ (with two explanatory variables) we get two nromal equations and solving the equations we can find out the values of $\hat{\beta}_1$ and $\hat{\beta}_2$.

Now $\hat{\beta} = (X'X)^{-1}X'Y$

$$= \frac{1}{|X'X|} \text{ adj } (X'X) \cdot X'Y$$

Now Adj $(X'X)$ = Transpose of matrix of co-factors of $(X'X)$. Now matrix of co-factors of $(X'X)$

$$= \begin{bmatrix} \Sigma x_{2i}^2 & -\Sigma x_{1i}x_{2i} \\ -\Sigma x_{1i}x_{2i} & \Sigma x_{1i}^2 \end{bmatrix} \text{ and } |X'X| = \Sigma x_{1i}^2 \Sigma x_{2i}^2 - [\Sigma x_{1i}x_{2i}]^2$$

$$\therefore \text{ Adj } (X'X) = \begin{bmatrix} \Sigma x_{2i}^2 & -\Sigma x_{1i}x_{2i} \\ -\Sigma x_{1i}x_{2i} & \Sigma x_{1i}^2 \end{bmatrix}$$

$$\therefore \hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \frac{1}{|X'X|} \begin{bmatrix} \Sigma x_{2i}^2 & -\Sigma x_{1i}x_{2i} \\ -\Sigma x_{1i}x_{2i} & \Sigma x_{1i}^2 \end{bmatrix} \begin{bmatrix} \Sigma x_{1i}y_i \\ \Sigma x_{2i}y_i \end{bmatrix}$$

$$\therefore \hat{\beta}_1 = \frac{1}{|X'X|}\left[\Sigma x_{2i}^2 \Sigma x_{1i}y_i - \Sigma x_{1i}x_{2i}\Sigma x_{2i}y_i\right] = \frac{\Sigma x_{2i}^2 \Sigma x_{1i}y_i - \Sigma x_{1i}x_{2i} \cdot \Sigma x_{2i}y_i}{\Sigma x_{1i}^2 \cdot \Sigma x_i^2 - (\Sigma x_{1i}x_{2i})^2}$$

$$\text{and } \hat{\beta}_2 = \frac{1}{|X'X|}[-\Sigma x_{1i}x_{2i} \cdot \Sigma x_{1i}y_i + \Sigma x_{1i}^2 \Sigma x_{2i}y_i] = \frac{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}y_i - \Sigma x_{1i}x_{2i}\Sigma x_{1i}y_i}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i}x_{2i})^2}$$

when $\hat{\beta}_1$ and $\hat{\beta}_2$ are known, $\hat{\beta}_0$ can be obtained from the relation

$$\overline{Y} = \hat{\beta}_0 + \hat{\beta}_1\overline{X}_1 + \hat{\beta}_2\overline{X}_2 \quad \therefore \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}_1 - \hat{\beta}_2\overline{X}_2$$

We can also find out the values of $\hat{\beta}_1$ and $\hat{\beta}_2$ directly by using Cramer's rule :

Since $\quad \hat{\beta} = (X'X)^{-1}X'Y$ or, $(X'X)\hat{\beta} = X'Y$

or, $$\begin{bmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i}x_{2i} \\ \Sigma x_{1i}x_{2i} & \Sigma x_{2i}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

or, $$\begin{bmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i}x_{2i} \\ \Sigma x_{1i}x_{2i} & \Sigma x_{2i}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \Sigma x_{1i}y_i \\ \Sigma x_{2i}y_i \end{bmatrix}$$

or, $\hat{\beta}_1 \Sigma x_{1i}^2 + \hat{\beta}_2 \Sigma x_{1i}x_{2i} = \Sigma x_{1i}y_i$ .... (A)

and $\hat{\beta}_1 \Sigma x_{1i}x_{2i} + \hat{\beta}_2 \Sigma x_{2i}^2 = \Sigma x_{2i}y_i$ .... (B)

Solving equations (A) and (B) by Cramer's rule
we have,

$$\hat{\beta}_1 = \frac{\begin{vmatrix} \Sigma x_{1i}y_i & \Sigma x_{1i}x_{2i} \\ \Sigma x_{2i}y_i & \Sigma x_{2i}^2 \end{vmatrix}}{\begin{vmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i}x_{2i} \\ \Sigma x_{1i}x_{2i} & \Sigma x_{2i}^2 \end{vmatrix}} = \frac{\Sigma x_{2i}^2 \cdot \Sigma x_{1i}y_i - \Sigma x_{1i}x_{2i} \cdot \Sigma x_{2i}y_i}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i}x_{2i})^2}$$

and $$\hat{\beta}_2 = \frac{\begin{vmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i}y_i \\ \Sigma x_{1i}x_{2i} & \Sigma x_{2i}y_i \end{vmatrix}}{\begin{vmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i}x_{2i} \\ \Sigma x_{1i}x_{2i} & \Sigma x_{2i}^2 \end{vmatrix}} = \frac{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}y_i - \Sigma x_{1i}x_{2i} \cdot \Sigma x_{1i}y_i}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i}x_{2i})^2}$$

When $\hat{\beta}_1$ and $\hat{\beta}_2$ are known $\hat{\beta}_0$ is obtained from the relation

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \hat{\beta}_2 \bar{X}_2 \quad \therefore \; \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$$

Note : For the three variable linear regression equation

$[Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \; i = 1, 2, \ldots, n$ where $u_i \sim N(0, \sigma_u^2)]$

we can also find out the values of the regression parameters in another way. This method is described below :

The estimated regression line is given by

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} \text{ and } \bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \hat{\beta}_2 \bar{X}_2$$

where $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ are the OLS estimators of β_0, β_1 and β_2.

We obtain by subtraction, $\hat{y}_i = \hat{Y}_i - \bar{Y}$

$$= \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} - \hat{\beta}_0 - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$$

or, $\hat{y}_i = \hat{\beta}_1(X_{1i} - \bar{X}_1) + \hat{\beta}_2(X_{2i} - \bar{X}_2)$

or, $\hat{y}_i = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$ where $x_{1i} = X_{1i} - \bar{X}_1$, $x_{2i} = X_{2i} - \bar{X}_2$, $y_i = Y_i - \bar{Y}$

Now errors of estimate, $e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i})$

and $\Sigma e_i^2 = \Sigma(y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2$

The first order conditions for minimization require

$$\frac{\delta \Sigma e_i^2}{\partial \hat{\beta}_1} = 2\Sigma(y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})(-x_{1i}) = 0$$

or, $\Sigma x_{1i} y_i = \hat{\beta}_1 \Sigma x_{1i}^2 + \hat{\beta}_2 \Sigma x_{1i} x_{2i}$ ...... (1)

and $$\frac{\partial \Sigma e_i^2}{\partial \hat{\beta}_2} = 2\Sigma(y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})(-x_{2i}) = 0$$

or, $\Sigma x_{2i} y_i = \hat{\beta}_1 \Sigma x_{1i} x_{2i} + \hat{\beta}_2 \Sigma x_{2i}^2$ .... (2)

Now solving equations (1) and (2) by Cramer's rule we have,

$$\hat{\beta}_1 = \frac{\begin{vmatrix} \Sigma x_{1i} y_i & \Sigma x_{1i} x_{2i} \\ \Sigma x_{2i} y_i & \Sigma x_{2i}^2 \end{vmatrix}}{\begin{vmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i} x_{2i} \\ \Sigma x_{1i} x_{2i} & \Sigma x_{2i}^2 \end{vmatrix}} = \frac{\Sigma x_{2i}^2 \cdot \Sigma x_{1i} y_i - \Sigma x_{1i} x_{2i} \cdot \Sigma x_{2i} y_i}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2}$$

and $$\hat{\beta}_2 = \frac{\begin{vmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i} y_i \\ \Sigma x_{1i} x_{2i} & \Sigma x_{2i} y_i \end{vmatrix}}{\begin{vmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i} x_{2i} \\ \Sigma x_{1i} x_{2i} & \Sigma x_{2i}^2 \end{vmatrix}} = \frac{\Sigma x_{1i}^2 \cdot \Sigma x_{2i} y_i - \Sigma x_{1i} x_{2i} \cdot \Sigma x_{1i} y_i}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2}$$

When $\hat{\beta}_1$ and $\hat{\beta}_2$ are known, $\hat{\beta}_0$ can be obtained from the relation

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \hat{\beta}_2 \bar{X}_2 \quad \text{i.e., } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$$

Example 3.1. Consider the following regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ where u_i is normally distributed with mean 0 and variance σ_u^2

Y :	4	7	3	9	7
X_1 :	2	3	1	5	9
X_2 :	5	3	2	1	7

Estimate $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ (the OLS estimators of β_0, β_1 and β_2).

Solution : Calculations for the estimators of the regression parameters

Table 3.1

	Y_i	X_{1i}	X_{2i}	x_{1i}	x_{1i}^2	x_{2i}	x_{2i}^2	y_i	y_i^2	$x_{1i}x_{2i}$	$x_{1i}y_i$	$x_{2i}y_i$
				$=X_{1i}-\bar{X}_1$		$=X_{2i}-\bar{X}_2$		$=Y_i-\bar{Y}$				
	4	2	5	−2	4	1.4	1.96	−2	4	−2.3	4	−2.8
	7	3	3	−1	1	−0.6	0.36	1	1	0.6	−1	−0.6
$n=5$	3	1	2	−3	9	−1.6	2.56	−3	9	4.8	9	4.8
	9	5	1	1	1	−2.6	6.76	3	9	−2.6	3	−7.8
	7	9	7	5	25	3.4	11.56	1	1	17.0	5	3.4
	ΣY_i $=30$	ΣX_{1i} $=20$	ΣX_{2i} $=18$	Σx_{1i} $=0$	ΣX_{1i}^2 $=40$	Σx_{2i} $=0$	Σx_{2i}^2 $=23.20$	Σy_i $=0$	Σy_i^2 $=24$	$\Sigma x_{1i}x_{2i}$ $=17$	$\Sigma x_{1i}y_i$ $=20$	$\Sigma x_{2i}y_i$ $=-3$

Here $n = 5$ as five sets of values of the variables are given.

Now $\bar{Y} = \frac{\Sigma Y_i}{n} = \frac{30}{5} = 6$, $\bar{X}_1 = \frac{\Sigma X_{1i}}{n} = \frac{20}{5} = 4$ and $\bar{X}_2 = \frac{\Sigma X_{2i}}{n} = \frac{18}{5} = 3.6$

and $x_{1i} = X_{1i} - \bar{X}_1$, $x_{2i} = X_{2i} - \bar{X}_2$ and $y_i = Y_i - \bar{Y}$

We know that the values of β_1 and β_2 in the regression equation

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ can be obtained by the OLS method. The OLS estimators of β_1 and β_2 are $\hat{\beta}_1$ and $\hat{\beta}_2$

$$\text{where } \hat{\beta}_1 = \frac{\Sigma x_{2i}^2 \cdot \Sigma x_{1i}y_i - \Sigma x_{1i}x_{2i} \cdot \Sigma x_{2i}y_i}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i}x_{2i})^2} = \frac{23.20 \times 20 - 17 \times -3}{40 \times 23.20 - (17)^2}$$

$$= \frac{464 + 51}{928 - 289} = \frac{515}{639} = 0.8059 = 0.806$$

$$\text{and } \hat{\beta}_2 = \frac{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}y_i - \Sigma x_{1i}x_{2i} \cdot \Sigma x_{1i}y_i}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i}x_{2i})^2} = \frac{40 \times -3 - 17 \times 20}{40 \times 23.20 - (17)^2}$$

$$= \frac{-120 - 340}{928 - 289} = \frac{-460}{639} = -0.7198 = -0.720$$

$\therefore$ $\hat{\beta}_1 = 0.806$ and $\hat{\beta}_2 = -0.720$.

When $\hat{\beta}_1$ and $\hat{\beta}_2$ are known $\hat{\beta}_0$ can be obtained from the relation

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \hat{\beta}_2 \bar{X}_2$$

$$\therefore \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 = 6 - 0.806 \times 4 - (-0.720) \times 3.6 = 5.368$$

$\therefore$ $\hat{\beta}_0 = 5.368$, $\hat{\beta}_1 = 0.806$ and $\hat{\beta}_2 = -0.720$

3.2.1. The Regression Coefficients Expressed in terms of Variances (SDs) and Coefficient of Correlations

In a three variable linear regression model [$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, $i = 1, 2, \ldots, n$)

we have $\hat{\beta}_1 = \dfrac{\Sigma x_{2i}^2 \cdot \Sigma x_{1i} y_i - \Sigma x_{1i} x_{2i} \cdot \Sigma x_{2i} y_i}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2}$

We have assumed that $x_{1i} = X_{1i} - \bar{X}_1$, $x_{2i} = X_{2i} - \bar{X}_2$, $y_i = Y_i - \bar{Y}$

Now $\Sigma x_{1i}^2 = \Sigma (X_{1i} - \bar{X}_1)^2 = n\sigma_{X_1}^2$ where $\frac{1}{n}\Sigma (X_{1i} - \bar{X}_1)^2 = \sigma_{X_1}^2$

$\Sigma x_{2i}^2 = \Sigma (X_{2i} - \bar{X}_2)^2 = n\sigma_{X_2}^2$ where $\frac{1}{n}\Sigma (X_{2i} - \bar{X}_2)^2 = \sigma_{X_2}^2$

and $\Sigma x_{1i} x_{2i} = \Sigma (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2) = n \cdot \frac{1}{n} \Sigma (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)$

$$= n\operatorname{cov}(X_1, X_2) = nr_{X_1X_2} \cdot \sigma_{X_1}\sigma_{X_2}$$

since $r_{XY} = \dfrac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$

Similarly, $\Sigma x_{1i} y_i = \Sigma (X_{1i} - \bar{X}_1)(Y_i - \bar{Y})$

$$= n \cdot \frac{1}{n} \Sigma (X_{1i} - \bar{X}_1)(Y_i - \bar{Y}) = n\operatorname{cov}(X_1, Y) = nr_{X_1Y}\sigma_{X_1}\sigma_Y$$

and $\Sigma x_{2i} y_i = \Sigma (X_{2i} - \bar{X}_2)(Y_i - \bar{Y}) = n \cdot \frac{1}{n} \Sigma (X_{2i} - \bar{X}_2)(Y_i - \bar{Y})$

$$= n\operatorname{cov}(X_2, Y) = nr_{X_2Y} \cdot \sigma_{X_2} \cdot \sigma_Y$$

Now putting values in the expression of $\hat{\beta}_1$ we get

$$\hat{\beta}_1 = \frac{n\sigma_{X_2}^2 \cdot nr_{X_1Y} \cdot \sigma_{X_1}\sigma_Y - nr_{X_1X_2} \cdot \sigma_{X_1}\sigma_{X_2} \cdot nr_{X_2Y} \cdot \sigma_{X_2} \cdot \sigma_Y}{n\sigma_{X_1}^2 \cdot n\sigma_{X_2}^2 - (n \cdot r_{X_1X_2} \cdot \sigma_{X_1}\sigma_{X_2})^2}$$

$$= \frac{n^2 \cdot \sigma_{X_2}^2 \cdot r_{X_1Y} \cdot \sigma_{X_1}\sigma_Y - n^2 r_{X_1X_2}\sigma_{X_1}\sigma_{X_2}^2 \cdot r_{X_2Y} \cdot \sigma_Y}{n^2\sigma_{X_1}^2 \cdot \sigma_{X_2}^2 - n^2 r_{X_1X_2}^2 \cdot \sigma_{X_1}^2\sigma_{X_2}^2}$$

$$= \frac{\sigma_{X_2}^2 r_{X_1Y}\sigma_{X_1}\sigma_Y - r_{X_1X_2}\sigma_{X_1}\sigma_{X_2}^2 r_{X_2Y}\sigma_Y}{\sigma_{X_1}^2\sigma_{X_2}^2 - r_{X_1X_2}^2\sigma_{X_1}^2\sigma_{X_2}^2}$$

$$= \frac{\sigma_{X_2}^2\sigma_{X_1}\sigma_Y(r_{X_1Y} - r_{X_1X_2}r_{X_2Y})}{\sigma_{X_1}^2\sigma_{X_2}^2(1 - r_{X_1X_2}^2)} = \frac{\sigma_Y}{\sigma_{X_1}}\left[\frac{r_{X_1Y} - r_{X_1X_2} \cdot r_{X_2Y}}{1 - r_{X_1X_2}^2}\right]$$

$$\therefore \hat{\beta}_1 = \frac{\sigma_Y}{\sigma_{X_1}}\left[\frac{r_{X_1Y} - r_{X_1X_2} \cdot r_{X_2Y}}{1 - r_{X_1X_2}^2}\right]$$

Proceeding in the same way we can get

$$\hat{\beta}_2 = \frac{\sigma_Y}{\sigma_{X_2}}\left[\frac{r_{X_2Y} - r_{X_1X_2}\cdot r_{X_1Y}}{1-r^2_{X_1X_2}}\right]$$ [where the symbols have their usual meanings]

3.2.2. Determination of Variances and Covariances of the Estimators of the Regression Parameters in Three Variable Linear Regression Model

For a three variable linear regression model we assume the regression equation in the form :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i,\ i = 1, 2, \ldots, n$$

$$\therefore\ \bar{Y} = \beta_0 + \beta_1 \bar{X}_1 + \beta_2 \bar{X}_2 + \bar{u},\ \text{where } \bar{u} = 0$$

Now $Y_i - \bar{Y} = \beta_1 (X_{1i} - \bar{X}_1) + \beta_2 (X_{2i} - \bar{X}_2) + (u_i - \bar{u})$

or, $y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$ where $x_{1i} = X_{1i} - \bar{X}_1$, $x_{2i} = X_{2i} - \bar{X}_2$ and $\bar{u} = 0$

In vector-matrix form the set of n equations (for $i = 1, 2, \ldots, n$) can be written as

$$Y = X\beta + u$$

where $X = \begin{bmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ \vdots & \vdots \\ x_{1n} & x_{2n} \end{bmatrix}_{n\times 2}$, $Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1}$, $u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}_{n\times 1}$ and $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}_{2\times 1}$.

Now $$X'X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \end{bmatrix} \begin{bmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ \vdots & \vdots \\ x_{1n} & x_{2n} \end{bmatrix} = \begin{bmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i}x_{2i} \\ \Sigma x_{1i}x_{2i} & \Sigma x_{2i}^2 \end{bmatrix}$$

Now $$D(\hat{\beta}) = \begin{bmatrix} E(\hat{\beta}_1 - \beta_1)^2 & E(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2) \\ E(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2) & E(\hat{\beta}_2 - \beta_2)^2 \end{bmatrix}$$

since $E(\hat{\beta}_1) = \beta_1$ and $E(\hat{\beta}_2) = \beta_2$

$$= \begin{bmatrix} \text{var}(\hat{\beta}_1) & \text{cov}(\hat{\beta}_1, \hat{\beta}_2) \\ \text{cov}(\hat{\beta}_1, \hat{\beta}_2) & \text{var}(\hat{\beta}_2) \end{bmatrix} = \sigma_u^2 (X'X)^{-1}$$

[See Property 2 of OLS estimator vector $\hat{\beta}$]

$$= \sigma_u^2 \begin{bmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i}x_{2i} \\ \Sigma x_{1i}x_{2i} & \Sigma x_{2i}^2 \end{bmatrix}^{-1}$$

$$= \sigma_u^2 \frac{1}{\begin{vmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i}x_{2i} \\ \Sigma x_{1i}x_{2i} & \Sigma x_{2i}^2 \end{vmatrix}} \cdot \text{Adj} \begin{bmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i}x_{2i} \\ \Sigma x_{1i}x_{2i} & \Sigma x_{2i}^2 \end{bmatrix}$$

$$= \frac{\sigma_u^2}{\begin{vmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i}x_{2i} \\ \Sigma x_{1i}x_{2i} & \Sigma x_{2i}^2 \end{vmatrix}} \begin{bmatrix} \Sigma x_{2i}^2 & -\Sigma x_{1i}x_{2i} \\ -\Sigma x_{1i}x_{2i} & \Sigma x_{1i}^2 \end{bmatrix}$$

Now, $\begin{bmatrix} \text{var}(\hat{\beta}_1) & \text{cov}(\hat{\beta}_1, \hat{\beta}_2) \\ \text{cov}(\hat{\beta}_1, \hat{\beta}_2) & \text{var}(\hat{\beta}_2) \end{bmatrix} = \dfrac{\sigma_u^2}{\Sigma x_{1i}^2 \Sigma x_{2i}^2 - (\Sigma x_{1i}x_{2i})^2} \begin{bmatrix} \Sigma x_{2i}^2 & -\Sigma x_{1i}x_{2i} \\ -\Sigma x_{1i}x_{2i} & \Sigma x_{1i}^2 \end{bmatrix}$

$$= \sigma_u^2 \begin{bmatrix} \dfrac{\Sigma x_{2i}^2}{\Sigma x_{1i}^2 \Sigma x_{2i}^2 - (\Sigma x_{1i}x_{2i})^2} & \dfrac{-\Sigma x_{1i}x_{2i}}{\Sigma x_{1i}^2 \Sigma x_{2i}^2 - (\Sigma x_{1i}x_{2i})^2} \\ \dfrac{-\Sigma x_{1i}x_{2i}}{\Sigma x_{1i}^2 \Sigma x_{2i}^2 - (\Sigma x_{1i}x_{2i})^2} & \dfrac{\Sigma x_{1i}^2}{\Sigma x_{1i}^2 \Sigma x_{2i}^2 - (\Sigma x_{1i}x_{2i})^2} \end{bmatrix}$$

$$\therefore \text{var}(\hat{\beta}_1) = \frac{\sigma_u^2 \Sigma x_{2i}^2}{\Sigma x_{1i}^2 \Sigma x_{2i}^2 - (\Sigma x_{1i}x_{2i})^2} = \frac{\sigma_u^2 n \sigma_{X_2}^2}{n\sigma_{X_1}^2 \sigma_{X_2}^2 - n^2 r_{X_1X_2}^2 \sigma_{X_1}^2 \sigma_{X_2}^2}$$

$$\therefore \text{var}(\hat{\beta}_1) = \frac{\sigma_u^2 n \sigma_{X_2}^2}{n\sigma_{X_1}^2 n\sigma_{X_2}^2 (1 - r_{X_1X_2}^2)} = \frac{\sigma_u^2}{n\sigma_{X_1}^2 (1 - r_{X_1X_2}^2)}$$

[Since $\Sigma x_{2i}^2 = \Sigma(X_{2i} - \overline{X}_2)^2 = n\frac{1}{n}\Sigma(X_{2i} - \overline{X})^2 = n\sigma_{X_2}^2$, similarly, $\Sigma x_{1i}^2 = n\sigma^2 X_1$

and $\Sigma x_{1i}x_{2i} = \Sigma(X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2) = n \cdot \frac{1}{n}\Sigma(X_{1i} - \overline{X}_1)(X_{2i} - \overline{X}_2)$

$= n \text{ cov }(X_1, X_2) = n \cdot r_{X_1X_2} \cdot \sigma_{X_1}\sigma_{X_2}$ as $\dfrac{\text{cov}(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}} = r_{X_1X_2}$]

Similarly, $\text{var}(\hat{\beta}_2) = \dfrac{\sigma_u^2 \Sigma x_{1i}^2}{\Sigma x_{1i}^2 \Sigma x_{2i}^2 - (\Sigma x_{1i}x_{2i})^2}$

$$= \frac{\sigma_u^2 n\sigma_{X_1}^2}{n\sigma_{X_1}^2 n\sigma_{X_2}^2 - n^2 r_{X_1X_2}^2 \sigma_{X_1}^2 \sigma_{X_2}^2} = \frac{\sigma_u^2 n\sigma_{X_1}^2}{n\sigma_{X_1}^2 n\sigma_{X_2}^2 (1 - r_{X_1X_2}^2)}$$

$$\therefore \text{var}(\hat{\beta}_2) = \frac{\sigma_u^2}{n\sigma_{X_2}^2 (1 - r_{X_1X_2}^2)}$$

Again, $\text{cov}(\hat{\beta}_1, \hat{\beta}_2) = \dfrac{-\sigma_u^2 \Sigma x_{1i}x_{2i}}{\Sigma x_{1i}^2 \Sigma x_{2i}^2 - (\Sigma x_{1i}x_{2i})^2}$

$$= \frac{-\sigma_u^2 n r_{X_1X_2} \sigma_{X_1} \sigma_{X_2}}{n\sigma_{X_1}^2 n\sigma_{X_2}^2 - n^2 r_{X_1X_2}^2 \sigma_{X_1}^2 \sigma_{X_2}^2} = \frac{-n\sigma_u^2 r_{X_1X_2} \sigma_{X_1} \sigma_{X_2}}{n^2 \sigma_{X_1}^2 \sigma_{X_2}^2 (1 - r_{X_1X_2}^2)}$$

$$\therefore \operatorname{cov}(\hat{\beta}_1,\hat{\beta}_2)=\frac{-\sigma_u^2 r_{X_1X_2}}{n\sigma_{X_1}\sigma_{X_2}(1-r_{X_1X_2}^2)}$$

Now, $\operatorname{var}(\hat{\beta}_1+\hat{\beta}_2)=\operatorname{var}(\hat{\beta}_1)+\operatorname{var}(\hat{\beta}_2)+2\operatorname{cov}(\hat{\beta}_1,\hat{\beta}_2)$

$$=\frac{\sigma_u^2}{n\sigma_{X_1}^2(1-r_{X_1X_2}^2)}+\frac{\sigma_u^2}{n\sigma_{X_2}^2(1-r_{X_1X_2}^2)}-\frac{2\sigma_u^2 r_{X_1X_2}}{n\sigma_{X_1}\sigma_{X_2}(1-r_{X_1X_2}^2)}$$

and $\operatorname{var}(\hat{\beta}_1-\hat{\beta}_2)=\operatorname{var}(\hat{\beta}_1)+\operatorname{var}(\hat{\beta}_2)-2\operatorname{cov}(\hat{\beta}_1,\hat{\beta}_2)$

$$=\frac{\sigma_u^2}{n\sigma_{X_1}^2(1-r_{X_1X_2}^2)}+\frac{\sigma_u^2}{n\sigma_{X_2}^2(1-r_{X_1X_2}^2)}+\frac{2\sigma_u^2 r_{X_1X_2}}{n\sigma_{X_1}\sigma_{X_2}(1-r_{X_1X_2}^2)}$$

since $\bar{Y}=\hat{\beta}_0+\hat{\beta}_1\bar{X}_1+\hat{\beta}_2\bar{X}_2 \therefore \hat{\beta}_0=\bar{Y}-\hat{\beta}_1\bar{X}_1-\hat{\beta}_2\bar{X}_2$

It can be seen that

$$\operatorname{var}(\hat{\beta}_0)=\frac{\sigma_u^2}{n}+\bar{X}_1^2\operatorname{var}(\hat{\beta}_1)+2\bar{X}_1\bar{X}_2\operatorname{cov}(\hat{\beta}_1,\hat{\beta}_2)+\bar{X}_2^2\operatorname{var}(\hat{\beta}_2)$$

$$\operatorname{cov}(\hat{\beta}_0,\hat{\beta}_1)=-[\bar{X}_1\operatorname{var}(\hat{\beta}_1)+\bar{X}_2\operatorname{cov}(\hat{\beta}_1,\hat{\beta}_2)]$$

and $\operatorname{cov}(\hat{\beta}_0,\hat{\beta}_2)=-[\bar{X}_1\operatorname{cov}(\hat{\beta}_1,\hat{\beta}_2)+\bar{X}_2\operatorname{var}(\hat{\beta}_2)]$

Note : In calculating $\operatorname{var}(\hat{\beta}_0)$, $\operatorname{var}(\hat{\beta}_1)$, $\operatorname{var}(\hat{\beta}_2)$, $\operatorname{cov}(\hat{\beta}_1,\hat{\beta}_2)$, $\operatorname{cov}(\hat{\beta}_0,\hat{\beta}_1)$ and $\operatorname{cov}(\hat{\beta}_0,\hat{\beta}_2)$ if σ_u^2 is not known it is to be replaced by its unbiased estimator $\hat{\sigma}_u^2=\Sigma e_i^2/(n-3)$.

Example 3.2. The following table presents data on a sample of five persons randomly drawn from a large firm giving their annual salaries in thousands of dollars (Y), years of education (X_1) and years of experience with the firm they are working for (X_2) :

Y :	30	20	36	24	40
X_1 :	4	3	6	4	8
X_2 :	10	8	11	9	12

Assuming a linear regression of the form

$$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+u_i,\quad u_i\sim N(0,\sigma_u^2),$$

(i) find the OLS estimators : $\hat{\beta}_0,\hat{\beta}_1$ and $\hat{\beta}_2$.

(ii) find the value of $r_{X_1X_2}$.

(iii) find the estimated regression equation.

(iv) find Σe_i^2.

(v) find the values of $\operatorname{var}(\hat{\beta}_1)$, $\operatorname{var}(\hat{\beta}_2)$ and $\operatorname{var}(\hat{\beta}_0)$

(vi) find the value of $\operatorname{cov}(\hat{\beta}_1,\hat{\beta}_2)$

Solution :

Calculation Table 3.2

Y_i	X_{1i}	X_{2i}	y_i	y_i^2	x_{1i}	x_{1i}^2	x_{2i}	x_{2i}^2	$x_{1i}y_i$	$x_{2i}y_i$	$x_{1i}x_{2i}$	e_i	e_i^2
			$= Y_i - \overline{Y}$		$= X_{1i} - \overline{X}_1$		$= X_{2i} - \overline{X}_2$				$= Y_i - \hat{Y}_i$		
30	4	10	0	0	–1	1	0	0	0	0	0	–0.25	0.0625
20	3	8	–10	100	–2	4	–2	4	20	20	4	0.5	0.2500
36	6	11	6	36	1	1	1	1	6	6	1	0.75	0.5625
24	4	9	–6	36	–1	1	–1	1	6	6	1	–0.75	0.5625
40	8	12	10	100	3	9	2	4	30	20	6	–0.25	0.0625
ΣY_i $= 150$	ΣX_{1i} $= 25$	ΣX_{2i} $= 50$	Σy_i $= 0$	Σy_i^2 $= 272$	Σx_{1i} $= 0$	Σx_{1i}^2 $= 16$	Σx_{2i} $= 0$	Σx_{2i}^2 $= 10$	$\Sigma x_{1i}y_i$ $= 62$	$\Sigma x_{2i}y_i$ $= 52$	$\Sigma x_{1i}x_{2i}$ $= 12$	Σe_i $= 0$	Σe_i^2 $= 1.5$

Here $n = 5$ as five sets of values are given.

Now $\overline{Y} = \dfrac{\Sigma Y_i}{n} = \dfrac{150}{5} = 30$, $\overline{X}_1 = \dfrac{\Sigma X_{1i}}{n} = \dfrac{25}{5} = 5$, $\overline{X}_2 = \dfrac{\Sigma X_{2i}}{n} = \dfrac{50}{5} = 10$

(i) We have to find out the OLS estimators $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$.

We know that $\hat{\beta}_1 = \dfrac{\Sigma x_{2i}^2 \cdot \Sigma x_{1i} y_i - \Sigma x_{1i} x_{2i} \Sigma x_{2i} y_i}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2}$

We now put the values from the calculation table and get

$$\hat{\beta}_1 = \frac{10 \times 62 - 12 \times 52}{16 \times 10 - (12)^2} = \frac{620 - 624}{160 - 144} = \frac{-4}{16} = -0.25$$

Similarly, $\hat{\beta}_2 = \dfrac{\Sigma x_{1i}^2 \cdot \Sigma x_{2i} y_i - \Sigma x_{1i} x_{2i} \cdot \Sigma x_{1i} y_i}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2} = \dfrac{16 \times 52 - 12 \times 62}{16 \times 10 - (12)^2} = \dfrac{88}{16} = 5.5 \therefore \hat{\beta}_2 = 5.5$

When $\hat{\beta}_1$ and $\hat{\beta}_2$ are known, $\hat{\beta}_0$ can be obtained from the relation

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 = 30 - (-0.25) \times 5 - 5.5 \times 10$$

$$\therefore \hat{\beta}_0 = 30 + 1.25 - 55 = -23.75$$

Thus the OLS estimators of the parameters are $\hat{\beta}_0 = -23.75$, $\hat{\beta}_1 = -0.25$ and $\hat{\beta}_2 = 5.5$

(ii) We have to find out the value of product moment correlation coefficient between two explanatory variables X_1 and X_2 i.e., $r_{X_1 X_2}$.

$$\text{We know that } r_{X_1 X_2} = \frac{\text{cov}(X_1, X_2)}{\sigma_{X_1} \cdot \sigma_{X_2}} = \frac{\frac{1}{n}\Sigma(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sqrt{\frac{1}{n}\Sigma(X_{1i} - \bar{X}_1)^2}\sqrt{\frac{1}{n}\Sigma(X_{2i} - \bar{X}_2)^2}}$$

$$= \frac{\frac{1}{n}\Sigma x_{1i} x_{2i}}{\sqrt{\frac{\Sigma x_{1i}^2}{n}} \cdot \sqrt{\frac{\Sigma x_{2i}^2}{n}}} = \frac{\Sigma x_{1i} x_{2i}}{\sqrt{\Sigma x_{1i}^2} \cdot \sqrt{\Sigma x_{2i}^2}} = \frac{12}{\sqrt{16 \times 10}} = 0.9486 = 0.95 \therefore r_{X_1 X_2} = 0.95$$

(iii) The estimated regression line is given by

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$$

$\therefore \hat{Y}_i = -23.75 - 0.25\, X_{1i} + 5.5\, X_{2i}$ is the estimated regression line/equation.

(iv) We have to find out the value of Σe_i^2

where $\Sigma e_i^2 = \Sigma(Y_i - \hat{Y}_i)^2 = (Y_1 - \hat{Y}_1)^2 + (Y_2 - \hat{Y}_2)^2 + (Y_3 - \hat{Y}_3)^2$

$$+ (Y_4 - \hat{Y}_4)^2 + (Y_5 - \hat{Y}_5)^2$$

$= (-0.25)^2 + (0.5)^2 + (0.75)^2 + (-0.75)^2 + (-0.25)^2 = 1.5 \quad \therefore \Sigma e_i^2 = 1.5.$

[In particular when $Y = Y_1 = 30$, $X_{1i} = 4$, $X_{21} = 10$ then

$\hat{Y}_i = \hat{Y}_1 = -23.75 - 0.25 \times 4 + 5.5 \times 10 = 30.25$

$\therefore e_1 = Y_1 - \hat{Y}_1 = 30 - 30.25 = -0.25.$

Similarly other e_i values are obtained as follows :

$\hat{Y}_i = -23.75 - 0.25X_{1i} + 5.5X_{2i}$

when $Y = Y_2 = 20, X_{12} = 3, X_{22} = 8$

then $\hat{Y}_i = \hat{Y}_2 = -23.75 - 0.25 \times 3 + 5.5 \times 8$

$= -23.75 - 0.75 + 44 = 19.5$

$\therefore e_2 = Y_2 - \hat{Y}_2 = 20 - 19.5 = 0.5$

When $Y = Y_3 = 36, X_{13} = 6, X_{23} = 11,$

then $\hat{Y}_i = \hat{Y}_3 = -23.75 - 0.75 \times 6 + 5.5 \times 11 = -23.75 - 1.5 + 60.5 = 35.25$

$\therefore e_3 = Y_3 - \hat{Y}_3 = 36 - 35.25 = 0.75$

When $Y = Y_4 = 24, X_{14} = 4, X_{24} = 9,$

then $\hat{Y}_i = \hat{Y}_4 = -23.75 - 0.25 \times 4 + 5.5 \times 9 = -23.75 - 1 + 49.5 = 24.75$

$\therefore e_4 = Y_4 - \hat{Y}_4 = 24 - 24.75 = -0.75$

When $Y = Y_5 = 40, X_{15} = 8, X_{25} = 12,$

then $\hat{Y}_i = \hat{Y}_5 = -23.75 - 0.25 \times 8 + 5.5 \times 12$

$= -23.75 - 2 + 66 = 40.25$

$\therefore e_5 = Y_5 - \hat{Y}_5 = 40 - 40.25 = -0.25$

(v) Now we have to calculate the variances of the OLS estimators of the regression parameters, $\text{var}(\hat{\beta}_1)$, $\text{var}(\hat{\beta}_2)$ and $\text{var}(\hat{\beta}_0)$.

We know that $\text{var}(\hat{\beta}_1) = \dfrac{\sigma_u^2 \Sigma x_{2i}^2}{\Sigma x_{1i}^2 \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2} = \dfrac{\left(\dfrac{\Sigma e_i^2}{n-3}\right)\Sigma x_{2i}^2}{\Sigma x_{1i}^2 \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2}$. Here σ_u^2 is not known and hence it is replaced by its unbiased estimator $\hat{\sigma}_u^2 = \dfrac{\Sigma e_i^2}{n-3}$.

Here $\dfrac{\Sigma e_i^2}{n-3} = \dfrac{1.5}{5-3} = \dfrac{1.5}{2} = 0.75$

$$\therefore \text{var}(\hat{\beta}_1) = \frac{\left(\dfrac{\Sigma e_i^2}{n-3}\right)\Sigma x_{2i}^2}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_i x_{2i})^2} = \frac{0.75 \times 10}{16 \times 10 - (12)^2}$$

$= \dfrac{7.5}{16} = 0.4687 \quad \therefore \text{var}(\hat{\beta}_1) = 0.4687$

$$\text{Similarly, } \operatorname{var}(\hat{\beta}_2) = \frac{\left(\frac{\Sigma e_i^2}{n-3}\right)\Sigma x_{1i}^2}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i}x_{2i})^2} = \frac{0.75 \times 16}{16 \times 10 - (12)^2} = \frac{12}{16} = 0.75$$

$\therefore \operatorname{var}(\hat{\beta}_2) = 0.75$

and $\operatorname{var}(\hat{\beta}_0) = \frac{\sigma_u^2}{n} + \bar{X}_1^2 \operatorname{var}(\hat{\beta}_1) + 2\bar{X}_1\bar{X}_2 \operatorname{cov}(\hat{\beta}_1, \hat{\beta}_2) + \bar{X}_0 \operatorname{var}(\hat{\beta}_2)$

We put $\sigma_u^2 = \hat{\sigma}_u^2 = \frac{\Sigma e_i^2}{n-3} = 0.75$, $n = 5$, $\bar{X}_1 = 5$, $\bar{X}_2 = 10$,

$\operatorname{var}(\hat{\beta}_1) = 0.4687$, $\operatorname{var}(\hat{\beta}_2) = 0.75$ and $\operatorname{cov}(\hat{\beta}_1, \hat{\beta}_2) = -0.5625$ (obtained from vi.)

Thus, we have $\operatorname{var}(\hat{\beta}_0) = \frac{0.75}{5} + (5)^2 \times 0.4687 + 2 \times 5 \times 10 \times -0.5625$

$$+ (10)^2 \times 0.75 = 30.6175$$

$\therefore \operatorname{var}(\hat{\beta}_0) = 30.6175$

(vi) We have to find out the value of $\operatorname{cov}(\hat{\beta}_1, \hat{\beta}_2)$.

Now $\operatorname{cov}(\hat{\beta}_1, \hat{\beta}_2) = \frac{-\sigma_u^2 \Sigma x_{1i}x_{2i}}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i}x_{2i})^2}$. Here σ_u^2 is not known and hence it is replaced by its unbiased estimator $\hat{\sigma}_u^2 = \frac{\Sigma e_i^2}{n-3} = 0.75$

$$\therefore \operatorname{cov}(\hat{\beta}_1, \hat{\beta}_2) = \frac{-0.75 \times 12}{16 \times 10 - (12)^2} = \frac{-9}{16} = -0.5625$$

$\therefore \operatorname{cov}(\hat{\beta}_1, \hat{\beta}_2) = -0.5625$.

3.3. Properties of OLS Estimator Vector $\hat{\beta}$

Let $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_K X_{Ki} + u_i$ be a General Linear Regression model.

In vector-matrix form the model takes the form : $Y = X\beta + u$

$$\text{where } Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n \times 1}, X = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{K1} \\ 1 & X_{12} & X_{22} & \cdots & X_{K2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{Kn} \end{bmatrix}_{n \times (K+1)},$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}_{(K+1) \times 1} \quad \text{and } u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}_{n \times 1}$$

It has been established earlier that $\hat{\beta}$ (OLS estimator vector of β) $=(X'X)^{-1}X'Y$. The OLS estimator vector $\hat{\beta}$ satisfies the following properties :

Property 1 : $\hat{\beta}$ is an unbiased estimator of β.

Proof : We know that $\hat{\beta}=(X'X)^{-1}X'Y$ where $Y = X\beta + u$.

Now $\hat{\beta}=(X'X)^{-1}X'(X\beta+u)=(X'X)^{-1}X'X\beta+(X'X)^{-1}X'u$.

or, $\hat{\beta}=I_{(K+1)}\cdot\beta+(X'X)^{-1}X'u=\beta+(X'X)^{-1}X'u$

where I_{K+1} = Identity matrix of order $(K + 1) \times (K + 1)$ having the same rank of the matrix X as well as $X'X$.

Now, $\hat{\beta}=\beta+(X'X)^{-1}X'u$. *This is a linear form in terms of vector matrix form.*

Again, $E(\hat{\beta}) = E[\beta+(X'X)^{-1}X'u]$

$$= E(\beta)+(X'X)^{-1}X'E(u) = E(\beta) = \beta \; [\because E\,(u) = 0]$$

$$\therefore E(\hat{\beta}) = \beta \text{ where } E\,(u) = E\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} E(u_1) \\ E(u_2) \\ \cdot \\ E(u_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0_{n\times 1} \text{ (Null vector)}$$

This shows that the OLS estimator of β is an unbiased estimator. *i.e.,* $E(\hat{\beta}) = \beta$

$$\text{Here } E(\hat{\beta}) = \begin{bmatrix} E(\hat{\beta}_0) \\ E(\hat{\beta}_1) \\ E(\hat{\beta}_2) \\ \vdots \\ E(\hat{\beta}_K) \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} \Rightarrow \begin{matrix} E(\hat{\beta}_0) = \beta_0 \\ E(\hat{\beta}_1) = \beta_1 \\ E(\hat{\beta}_2) = \beta_2 \\ \vdots \\ E(\hat{\beta}_K) = \beta_K \end{matrix}$$

This implies that OLS estimator of each parameter is an unbiased estimator. In terms of a linear regression model with two explanatory variables we have, $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, $i = 1, 2, ..., n$

$$\text{In this model we have, } \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}_{3\times 1}, \; \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}_{3\times 1}$$

$$\text{and } E(\hat{\beta}) = \begin{bmatrix} E(\hat{\beta}_0) \\ E(\hat{\beta}_1) \\ E(\hat{\beta}_{\beta 2}) \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}_{3\times 1}$$

$$\therefore E(\hat{\beta}) = \beta \text{ where } E(\hat{\beta}_0) = \beta_0, \; E(\hat{\beta}_1) = \beta_1, \; E(\hat{\beta}_2) = \beta_2$$

Property 2. The Dispersion matrix or variance-covariance matrix of $\hat{\beta}$ is given by $\sigma_u^2(X'X)^{-1}$

Proof : By definition, dispersion matrix or Varinace-Covariance matrix of $\hat{\beta}$ is given by $D(\hat{\beta}) = E[\hat{\beta}-\beta][\hat{\beta}-\beta]'$ where $E(\hat{\beta}) = \beta$

$$= \begin{bmatrix} E(\hat{\beta}_0-\beta_0)^2 & E(\hat{\beta}_1-\beta_1)(\hat{\beta}_0-\beta_0) & \cdots & E(\hat{\beta}_K-\beta_K)(\hat{\beta}_0-\beta_0) \\ E(\hat{\beta}_0-\beta_0)(\hat{\beta}_1-\beta_1) & E(\hat{\beta}_1-\beta_1)^2 & \cdots & E(\hat{\beta}_K-\beta_K)(\hat{\beta}_1-\beta_1) \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ E(\hat{\beta}_0-\beta_0)(\hat{\beta}_K-\beta_K) & E(\hat{\beta}_1-\beta_1)(\hat{\beta}_K-\beta_K) & \cdots & E(\hat{\beta}_K-\beta_K)^2 \end{bmatrix}$$

$$= \begin{bmatrix} \text{var}(\hat{\beta}_0) & \text{cov}(\hat{\beta}_0,\hat{\beta}_1) & \cdots & \text{cov}(\hat{\beta}_K,\hat{\beta}_0) \\ \text{cov}(\hat{\beta}_0,\hat{\beta}_1) & \text{var}(\hat{\beta}_1) & \cdots & \text{cov}(\hat{\beta}_K,\hat{\beta}_1) \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \text{cov}(\hat{\beta}_0,\hat{\beta}_K) & \text{cov}(\hat{\beta}_1,\hat{\beta}_K) & \cdots & \text{var}(\hat{\beta}_K) \end{bmatrix}$$

Here the diagonal terms are variances and non-diagonal terms are covariances. So, it is also called variance covariance matrix.

$$\therefore D(\hat{\beta}) = E[\hat{\beta}-\beta][\hat{\beta}-\beta]'$$
$$= E\left[(X'X)^{-1}X'u\right]\left[(X'X)^{-1}X'u\right]',$$
$$= E\left[(X'X)^{-1}X'uu'X(X'X)^{-1}\right]$$
$$= (X'X)^{-1}X'E(uu')X(X'X)^{-1}$$
$$= (X'X)^{-1}X'\sigma_u^2 I_n \cdot X(X'X)^{-1}$$
$$= \sigma_u^2(X'X)^{-1}(X'X)(X'X)^{-1}$$
$$= \sigma_u^2(X'X)^{-1}I_{K+1} = \sigma_u^2(X'X)^{-1}$$
$$\therefore D(\hat{\beta}) = \sigma_u^2(X'X)^{-1}$$

Since $\hat{\beta} = \beta + (X'X)^{-1}X'u$

$\therefore \hat{\beta}-\beta = (X'X)^{-1}X'u$

Since $D(u)$

$= E[u-E(u)][u-E(u)]'$

$= E(uu')$ as $E(u) = 0$

$= \sigma_u^2 I_n$ [See 3.1 (ii)]

where I_{K+1} = Identity matrix of order $(K+1) \times (K+1)$

Proceeding in the same way we can also derive the result $D(\hat{\beta}) = \sigma_u^2(X'X)^{-1}$ for a regression model with two explanatory variables

i.e., $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, $i = 1, 2, 3, \ldots, n$

Property 3. *j*th element of $\hat{\beta}$ is the best linear unbiased estimator of the *j*th element of β. Alternatively, $\hat{\beta}$ is the Best Linear Unbiased Estimator (BLUE) of β

Proof : Since, $\hat{\beta} = (X'X)^{-1}X'Y$

Let us define $\hat{\beta}_0 = e_1' \hat{\beta}$ where $e_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{(K+1)\times 1}$ = First unit vector

$\therefore e_1' = [1, 0, \ldots, 0]_{1\times(K+1)}$

So, $\hat{\beta}_0 = e_1' \hat{\beta} = [1 \quad 0 \quad 0 \quad \ldots \quad 0] \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_K \end{bmatrix}$

Here $\hat{\beta}_0 = e_1'(X'X)^{-1} X'Y$.

Let us define $\beta_0^* = C'Y$ where $C = \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_n \end{bmatrix}_{n\times 1}$, an arbitrarily chosen column vector.

Now we have to find out the conditions under which β_0^* is an unbiased estimator of β_0.

Now, $\beta_0^* = C'Y = C'(X\beta + u)$ since $Y = X\beta + u$

$$= C'X\beta + C'u$$

$$\therefore E(\beta_0^*) = E[C'X\beta] + E(C'u)$$

$$= C'X\beta + 0 = C'X\beta \quad [\because E(u) = 0_{n\times 1}]$$

$$\therefore E(\beta_0^*) = C'X\beta$$

Now $E(\beta_0^*) = \beta_0$ if $C'X = e_1'$

as $e_1' \beta = [\ 1 \quad 0 \quad \ldots \quad 0] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} = \beta_0$

This shows that β_0^* is an unbiased estimator of β_0. The condition for β_0^* to be an unbiased estimator of β_0 is given by $C'X = e_1'$.

Again, $\beta_0^* = C'X\beta + C'u$

$$= e_1'\beta + C'u = \beta_0 + [C_1 \quad C_2 \ \ldots . \ C_n] \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

or, $\beta_0^* = \beta_0 + \sum_{i=1}^{n} C_i u_i \;\therefore\; \beta_0^* - \beta_0 = \sum_{i=1}^{n} C_i u_i$

Now, $\text{var}(\beta_0^*) = E[\beta_0^* - \beta_0]^2 \;\because E(\beta_0^*) = \beta_0$

$$= E\left[\sum_{i=1}^{n} C_i u_i\right]^2 = \sigma_u^2 \sum_{i=1}^{n} C_i^2 \;\because E(u_i^2) = \sigma_u^2$$

$\therefore\; \text{var}(\beta_0^*) = \sigma_u^2 \sum_{i=1}^{n} C_i^2$

Now we have to minimise $\text{var}(\beta_0^*)$ subject to the condition that $E(\beta_0^*) = \beta_0$ through the choice of the vector C. In other words we have to minimise $\sigma_u^2 \sum_{i=1}^{n} C_i^2$ subject to the condition $C'X = e_1'$.

For the sake of simplicity we put $\sigma_u^2 = \frac{1}{2}$.

The Lagrangian is given by,

$$L = \frac{1}{2}\sum_{i=1}^{n} C_i^2 - [C'X - e_1']\lambda$$

where $\lambda = \begin{bmatrix} \lambda_0 \\ \lambda_1 \\ \vdots \\ \lambda_K \end{bmatrix}_{(K+1)\times 1}$, is the vector of Lagrangian multipliers.

$\therefore L = \frac{1}{2}C'C - [C'X - e_1']\lambda$

Now differentiating it with respect to C we get,

$\frac{\partial L}{\partial C} = C - X\lambda = 0_{n\times 1}$, a null column vector, where

$\frac{\partial}{\partial C}[C'C] = 2C$ and $\frac{\partial L}{\partial \lambda} = C'X - e_1' = 0_{1\times(K+1)}$, a null row vector.

$$\because C'C = [C_1\ C_2 \ldots C_n]\begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_n \end{bmatrix} = C_1^2 + C_2^2 + \cdots + C_n^2 = \sum_{i=1}^{n} C_i^2$$

$\therefore \frac{\partial L}{\partial C} = C - X\lambda = 0_{n\times 1}$

or, $C = X\lambda$ i.e., $C' = \lambda'X'$ and $C'X = \lambda'X'X$

Again, $\frac{\partial L}{\partial \lambda} = C'X - e_1' = 0_{1\times(K+1)} \;\therefore e_1' = C'X$

$\therefore e_1' = C'X = \lambda'X'X$ or, $\lambda'X'X = e_1'$

or, $\lambda' = e_1'(X'X)^{-1} \;\therefore C' = \lambda'X' = e_1'(X'X)^{-1}X'$

So, $\beta_0^* = C'Y = e_1'(X'X)^{-1}X'Y$

or, $\beta_0^* = e_1'\hat{\beta} = \hat{\beta}_0 \quad \left[\because \hat{\beta} = (X'X)^{-1}X'Y \quad \therefore e_1'\hat{\beta} = [1\ 0\ 0\ \ldots\ 0] \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_K \end{bmatrix} = \hat{\beta}_0 \right]$

So, it is proved that under the condition that β_0^* is an unbiased estimator of β_0, variance of β_0^* i.e. var (β_0^*) is minimum when $\beta_0^* = \hat{\beta}_0$.

Applying the similar mathematical technique it can be proved that

var (β_1^*) is minimum when $\beta_1^* = \hat{\beta}_1$

var (β_2^*) is minimum when $\beta_2^* = \hat{\beta}_2$

....

var (β_K^*) is minimum when $\beta_K^* = \hat{\beta}_K$

Hence it is proved that OLS estimators are the best linear unbiased estimators of the regression paramaters i.e., $\hat{\beta}$ is the BLUE of β. This is known as the GAUSS-MARKOV THEOREM.

In case of three variable Linear regression model ($Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, $i = 1, 2, \ldots, n$) we can prove that $\hat{\beta}$ is the BLUE of β where

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}_{3\times 1} \text{ and } \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}_{3\times 1}$$

Property 4 : Unbiased estimator of σ_u^2 is $\frac{e'e}{n-(K+1)} = \sum_{i=1}^{n} e_i^2 \Big/ \{n-(K+1)\}$ where K = number of explanatory variables and $(K + 1)$ = number of parameters including the constant intercept term.

Proof : Since $Y = X\beta + u$, $\hat{Y} = X\hat{\beta}$ and $Y = \hat{Y} + e$

or, $e = Y - \hat{Y}$ or, $e = Y - X\hat{\beta} \quad \therefore e' = (Y - X\hat{\beta})'$

Now $e'e = (Y - X\hat{\beta})'(Y - X\hat{\beta})$

$= [Y - X(X'X)^{-1}X'Y]'[Y - X(X'X)^{-1}X'Y]$ since $\hat{\beta} = (X'X)^{-1}X'Y$

$= [Y' - Y'X(X'X)^{-1}X'][Y - X(X'X)^{-1}X'Y]$

$= Y'[I - X(X'X)^{-1}X'][I - X(X'X)^{-1}X']Y$

$= Y'MMY$ where $M = I - X(X'X)^{-1}X'$

i.e., $e'e = Y'MMY = Y'MY$

$= (X\beta+u)'M(X\beta+u)$

$= (\beta'X'+u')M(X\beta+u)$

$= \beta'X'MX\beta+u'MX\beta+\beta'X'Mu+u'Mu$

I = Idenity matrix

where M is idempotent matrix

for which $M^2 = M$

Now if we put $M = I - X(X'X)^{-1}X'$ then

$\beta'X'MX\beta+u'MX\beta+\beta'X'Mu=0$

and hence $e'e = u'Mu \;=\; u'[I-X(X'X)^{-1}X']u$

$$= u'Iu - u'X(X'X)^{-1}X'u = \sum_{i=1}^{n} u_i^2 - u'X(X'X)^{-1}X'u$$

$\therefore E(e'e) = \Sigma E(u_i^2) - E[u'u \text{ trace } X(X'X)^{-1}X'] \left[\because E(u_i^2) = \sigma_u^2 \text{ and } u'u = \sum_{i=1}^{n} u_i^2\right]$

$= n\sigma_u^2 - \sigma_u^2\cdot(K+1)$ [Here trace $X(X'X)^{-1}X' = (K+1)$]

$\therefore E(e'e) = \sigma_u^2[n-(K+1)]$

or, $E\left[\dfrac{e'e}{n-(K+1)}\right] = \sigma_u^2$ or, $E\left[\dfrac{\Sigma e_i^2}{n-(K+1)}\right] = \sigma_u^2$

This proves that $\dfrac{e'e}{n-(K+1)} = \dfrac{\Sigma e_i^2}{n-(K+1)}$ is the unbiased estimator of σ_u^2.

In particular for a linear regression model with two explanatory variables,

$Y_i = \beta_0+\beta_1X_{1i}+\beta_2X_{2i}+u_i$,

we have $E\left[\dfrac{\Sigma e_i^2}{n-3}\right] = \sigma_u^2$ or, $E[\hat{\sigma}_u^2] = \sigma_u^2$ where $\hat{\sigma}_u^2 = \dfrac{\Sigma e_i^2}{n-3}$ and $K = 2$

3.4. MLE of β and σ_u^2 in the Multiple Regression Model

Let $Y_i = \beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\cdots+\beta_KX_{Ki}+u_i$ for $i = 1, 2, ..., n$, be the equation of the general linear regression model. In vector matrix form the set of n equations can be written as, $Y = X\beta+u$

where $Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1}$, $X = \begin{bmatrix} 1 & X_{11} & X_{21} & X_{K1} \\ 1 & X_{12} & X_{22} & X_{K2} \\ \vdots & & & \\ 1 & X_{1n} & X_{2n} & X_{Kn} \end{bmatrix}_{n\times(K+1)}$

$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}_{(K+1)\times 1}$, $u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}_{n\times 1}$

We are now interested in determining the (a) MLE of β and (b) MLE of σ_u^2 in this model.

Let us assume that each u_i is normally distributed with mean zero and variance σ_u^2 i.e., $u_i \sim N(0, \sigma_u^2)$. We also assume that the different disturbance terms $u_1, u_2, u_3, \ldots, u_n$ are independent.

Here $f\ (u_1, u_2, \ldots, u_n)$ is the joint p.d.f (probability density function) of the disturbance terms $(u_1, u_2, \ldots, u_n)$ and $f\ (u_i)$ is the p.d.f. of u_i for $i = 1, 2, \ldots, n$.

Since $u_1, u_2, \ldots, u_n$ are independent we can write $f(u_1, u_2, \ldots, u_n) = \prod\limits_{i=1}^{n} f_i(u_i)$.

Since $u_i \sim N(0, \sigma_u^2)$, then

$$f_i(u_i) = \frac{1}{\sqrt{2\pi} \cdot \sigma_u} e^{-\frac{1}{2}\frac{u_i^2}{\sigma_u^2}}, \text{ as } \bar{u} = 0$$

$$\text{So, } f\ (u_1, u_2, \ldots, u_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi} \cdot \sigma_u} e^{-\frac{1}{2}\frac{u_i^2}{\sigma_u^2}} = \frac{1}{(\sqrt{2\pi})^n \sigma_u^n} \cdot e^{-\frac{1}{2}\frac{\sum\limits_{i=1}^{n} u_i^2}{\sigma_u^2}}$$

This is the likelihood function of the parameters $\beta_0, \beta_1, \beta_3, \ldots, \beta_K$ and σ_u and is denoted by,

$$L(\beta', \sigma_u) = \frac{1}{(\sqrt{2\pi})^n \sigma_u^n} \cdot e^{-\frac{1}{2}\frac{\sum\limits_{i=1}^{n} u_i^2}{\sigma_u^2}} \quad \text{where } \beta' = [\beta_0, \beta_1, \ldots, \beta_K]_{(K+1)\times 1}$$

$$\text{Now } \log L = -\frac{n}{2}\log 2\pi - n\log\sigma_u - \frac{1}{2\sigma_u^2}\sum_{i=1}^{n} u_i^2$$

Now to obtain the MLE of the parameters Log L is to be maximised with respect to the parameters.

(a) To obtain the MLE of β we have to maximise Log L with respect to β which is equivalent to minimization of $\sum\limits_{i=1}^{n} u_i^2$ with respect to β.

So, we have to minimise Σu_i^2 through the choice of β.

$$\text{Since } u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}_{n\times 1}, \quad u' = [u_1, u_2, \ldots, u_n]_{1\times n}$$

$$\therefore u'u = \sum_{i=1}^{n} u_i^2$$

We know that $Y = X\beta + u$ or, $u = Y - X\beta$

$$\therefore \sum_{i=1}^{n} u_i^2 = u'u = (Y - X\beta)'(Y - X\beta) = (Y' - \beta'X')(Y - X\beta)$$

$$\therefore u'u = Y'Y - \beta'X'Y - Y'X\beta + \beta'X'X\beta$$

Now first order conditions of minimization will require

$$\frac{\partial(u'u)}{\partial\beta} = O_{(K+1)\times1} = \begin{bmatrix} \dfrac{\partial(u'u)}{\partial\beta_0} \\ \dfrac{\partial(u'u)}{\partial\beta_1} \\ \vdots \\ \dfrac{\partial(u'u)}{\partial\beta_K} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{(K+1)\times1} = O_{(K+1)\times1}$$

or, $\dfrac{\partial(u'u)}{\partial\beta} = -X'Y - X'Y + 2X'X\beta = 0$ -

$= -2X'Y + 2X'X\beta = 0$ [For details see the derivation of OLS of parameter vector β]

or, $-2X'Y + 2X'X\beta = 0$

or, $X'X\beta = X'Y$ or, $\beta = (X'X)^{-1}X'Y$

So, MLE of β is given by $(X'X)^{-1}X'Y$ which is identical with the OLS of β. This means that MLE of β equals OLS of β = $(X'X)^{-1}X'Y$.

(b) *We have to obtain the MLE of σ_u^2 :*

In order to derive the MLE of σ_u^2 we have to maximise

Log $L = -\dfrac{n}{2}\log 2\pi - n\log\sigma_u - \dfrac{1}{2\sigma_u^2}\sum_{i=1}^{n} u_i^2$ with respect to σ_u^2.

Now, Log $L = -\dfrac{n}{2}\log 2\pi - n\log\sigma_u - \dfrac{1}{2\sigma_u^2}u'u$ $\left[\because \sum_{i=1}^{n} u_i^2 = u'u\right]$

$= -\dfrac{n}{2}\log 2\pi - n\log\sigma_u - \dfrac{1}{2\sigma_u^2}[(Y - X\beta)'(Y - X\beta)]$ [Since $Y = X\beta + u$

$\therefore u = Y - X\beta$

and $u' = (Y - X\beta)'$]

Since it is proved that MLE of $\beta = \hat{\beta} = (X'X)^{-1}X'Y$ = OLS of β.

$\therefore \log L = -\dfrac{n}{2}\log 2\pi - n\log\sigma_u - \dfrac{1}{2\sigma_u^2}[(Y - X\hat{\beta})'(Y - X\hat{\beta})]$

$= -\dfrac{n}{2}\log 2\pi - n\log\sigma_u - \dfrac{1}{2\sigma_u^2}e'e$ [Since $\hat{Y} = X\hat{\beta}$, $e = Y - X\hat{\beta}$

and $e' = (Y - X\hat{\beta})'$ and $e'e = \sum_{i=1}^{n} e_i^2$]

or, $\log L = -\frac{n}{2}\log 2\pi - \left[n\log\sigma_u + \frac{1}{2\sigma_u^2}\sum_{i=1}^{n} e_i^2\right]$

Now $\log L$ is to be maximised with respect to σ_u or equivalently $n\log\sigma_u + \frac{1}{2\sigma_u^2}\sum_{i=1}^{n} e_i^2$ is to be minimised with respect to σ_u.

Now first order conditions of minimisation will require,

$$\frac{n}{\sigma_u} + \frac{1}{2}(-2)\frac{1}{\sigma_u^3}\sum_{i=1}^{n} e_i^2 = 0, \text{ a null vector}$$

or, $$\frac{n}{\sigma_u} - \left(\sum_{i=1}^{n} e_i^2 \Big/ \sigma_u^3\right) = 0 \text{ or, } \sigma_u^2 = \sum_{i=1}^{n} e_i^2 \Big/ n.$$

So, $\sum_{i=1}^{n} e_i^2 \Big/ n$ is the MLE of σ_u^2.

It should be noted that MLE of σ_u^2 is not an unbiased estimator of σ_u^2 but a consistent or asymptotically unbiased estimator of σ_u^2.

Since MLE of $\sigma_u^2 = \sum_{i=1}^{n} e_i^2 \Big/ n$

Now $$\sum_{i=1}^{n} e_i^2 \Big/ n = \frac{n-(K+1)}{n}\cdot\frac{\sum_{i=1}^{n} e_i^2}{n-(K+1)} = \left(1-\frac{K+1}{n}\right)\frac{\sum_{i=1}^{n} e_i^2}{n-(K+1)}$$

or, $$\sum_{i=1}^{n} e_i^2 \Big/ n = \left(1-\frac{K+1}{n}\right)\frac{\sum_{i=1}^{n} e_i^2}{n-(K+1)}$$

As $n \to \infty$, $\frac{(K+1)}{n} \to 0$ and hence

$\sum_{i=1}^{n} e_i^2 \Big/ n \to \sum_{i=1}^{n} e_i^2 \Big/ \{n-(K+1)\}$ where $\frac{\sum_{i=1}^{n} e_i^2}{n-(K+1)}$ is an unbiased estimator of σ_u^2.

This proves that MLE of σ_u^2 i.e. $\sum_{i=1}^{n} e_i^2 \Big/ n$ is an asymptotically unbiased or consistent estimator of σ_u^2. In a three variable (with two explanatory variables, i.e., when $K = 2$) linear regression model we have MLE of $\sigma_u^2 = \sum_{i=1}^{n} e_i^2 \Big/ n$ and unbiased estimator of $\sigma_u^2 = \sum_{i=1}^{n} e_i^2 \Big/ (n-3)$.

3.5. Expression of Multiple Correlation Coefficient in the General Linear Regression Model

Let $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_K X_{Ki} + u_i$ be the regression equation where $i = 1, 2, ..., n$.

Here Y is regressed on $X_1, X_2, X_3, ..., X_K$. So the multiple correlation coefficient is denoted by the symbol,

$$R^2_{Y \cdot X_1, X_2, ..., X_K} = \frac{\text{var}(\hat{Y})}{\text{var}(Y)} = \frac{\sum\limits_i (\hat{Y}_i - \bar{Y})^2 \Big/ n}{\sum\limits_i (Y_i - \bar{Y})^2 \Big/ n}$$

$$= \frac{\sum\limits_i \hat{Y}_i^2 - n\bar{Y}^2}{\sum\limits_i Y_i^2 - n\bar{Y}^2} = \frac{\hat{Y}'\hat{Y} - n\bar{Y}^2}{Y'Y - n\bar{Y}^2}$$

$$= \frac{(X\hat{\beta})'(X\hat{\beta}) - n\bar{Y}^2}{Y'Y - n\bar{Y}^2}$$

$$= \frac{\hat{\beta}'X'X\hat{\beta} - n\bar{Y}^2}{Y'Y - n\bar{Y}^2} = \frac{\hat{\beta}'X'Y - n\bar{Y}^2}{Y'Y - n\bar{Y}^2}$$

where $\hat{\beta} = (X'X)^{-1}X'Y \;\therefore\; (X'X)\hat{\beta} = X'Y$

Since $\hat{Y} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix}_{n \times 1}$

$\hat{Y}' = [\hat{Y}_1, \hat{Y}_2, ... \hat{Y}_n]_{1 \times n}$

where $\hat{Y} = X\hat{\beta}$ and $Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n \times 1}$

$Y' = [Y_1, Y_2, ..., Y_n]_{1 \times n}$

$\hat{\beta} = (X'X)^{-1}X'Y$

$\hat{\beta}' = Y'X(X'X)^{-1}$

3.6. The Multiple Coefficient of Determination R² and the Multiple Coefficient of Correlation in the Three-Variable Linear Regression Model

In the two variable case we have seen that R^2 (or r^2) measures the goodness of fit of the regression equation ($Y_i = \alpha + \beta X_i + u_i$, $i = 1, 2, ..., n$), that is it gives the proportion or percentage of the total variation in the dependent variable Y explained by the (single) explanatory variable X. This notation of R^2 can be easily extended to regression models containing more than two variables. Thus, in the three variable model we would like to know the proportion of the variation in Y explained by the variables X_1 and X_2 jointly.

This quantity that gives this information is known as the *multiple coefficient of determination* and is denoted by $R^2_{Y \cdot X_1, X_2}$ or simply R^2 and conceptually it is similer to r^2.

The estimated three variable regression line ($Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, i = 1, 2, ..., n$) is given by $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$ where $\hat{Y}_i$ is the estimated value of Y_i from the fitted regression line and is an estimator of true $E(Y_i / X_{1i}, X_{2i})$ and $\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \hat{\beta}_2 \bar{X}_2$.

Taking deviations from means we have

$$\hat{Y}_i - \bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} - \hat{\beta}_0 - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$$

or, $\hat{y}_i = \hat{\beta}_1(X_{1i} - \overline{X}_1) + \hat{\beta}_2(X_{2i} - \overline{X}_2)$

or, $\hat{y}_i = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$

where $\hat{y}_i = \hat{Y}_i - \overline{Y}$, $x_{1i} = X_{1i} - \overline{X}_1$ and $x_{2i} = X_{2i} - \overline{X}_2$

Now errors of estimate, $e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i})$

$\therefore y_i = \hat{y}_i + e_i$

Now $\Sigma y_i^2 = \Sigma \hat{y}_i^2 + \Sigma e_i^2 + 2\Sigma \hat{y}_i e_i$

or, $\Sigma y_i^2 = \Sigma \hat{y}_i^2 + \Sigma e_i^2 \quad \because \Sigma \hat{y}_i e_i = 0$

i.e., TSS = ESS + RSS

when TSS = Total Sum of Squares
ESS = Explained Sum of Squares
RSS = Residual Sum of Squares

Now by definition, $R^2_{Y \cdot X_1, X_2} = R^2 = \dfrac{\Sigma \hat{y}_i^2}{\Sigma y_i^2} = \dfrac{\text{ESS}}{\text{TSS}}$

$$= \frac{\Sigma(\hat{Y}_i - \overline{Y})^2}{\Sigma(Y_i - \overline{Y})^2} = \frac{\Sigma y_i^2 - \Sigma e_i^2}{\Sigma y_i^2}$$

$$= 1 - \frac{\Sigma e_i^2}{\Sigma y_i^2} = 1 - \frac{\text{RSS}}{\text{TSS}}.$$

Since
$\Sigma y_i^2 = \Sigma \hat{y}_i^2 + \Sigma e_i^2$
$\therefore \Sigma \hat{y}_i^2 = \Sigma y_i^2 - \Sigma e_i^2$

Since $e_i = y_i - \hat{y}_i \quad \therefore \Sigma e_i^2 = \Sigma e_i(y_i - \hat{y}_i)$

$= \Sigma e_i(y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})$ as $\hat{y}_i = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$

$= \Sigma e_i y_i - \hat{\beta}_1 \Sigma x_{1i} e_i - \hat{\beta}_2 \Sigma x_{2i} e_i$

$= \Sigma e_i y_i$

$= \Sigma (y_i - \hat{y}_i) y_i$

$= \Sigma y_i (y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})$

$= \Sigma y_i^2 - \hat{\beta}_1 \Sigma x_{1i} y_i - \hat{\beta}_2 \Sigma x_{2i} y_i$

Since $\Sigma x_i e_i = 0$, $\Sigma x_{2i} e_i = 0$

where $\Sigma x_{1i}(y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})$

$= \Sigma x_{1i} y_i - \hat{\beta}_1 \Sigma x_{1i}^2 - \hat{\beta}_2 \Sigma x_{1i} x_{2i}$

$= 0$

which follows from the first normal equation.

Similarly $\Sigma x_{2i} e_i = 0$

$\therefore \Sigma e_i^2 = \Sigma y_i^2 - \hat{\beta}_1 \Sigma x_{1i} y_i - \hat{\beta}_2 \Sigma x_{2i} y_i$

i.e., RSS = $\Sigma y_i^2 - \hat{\beta}_1 \Sigma x_{1i} y_i - \hat{\beta}_2 \Sigma x_{2i} y_i$

since Σy_i^2 = TSS

$\therefore$ ESS = $\hat{\beta}_1 \Sigma x_{1i} y_i + \hat{\beta}_2 \Sigma x_{2i} y_i$

$$\therefore R^2_{Y \cdot X_1, X_2} = \frac{\text{ESS}}{\text{TSS}} = \frac{\hat{\beta}_1 \Sigma x_{1i} y_i + \hat{\beta}_2 \Sigma x_{2i} y_i}{\Sigma y_i^2}$$

Example 3.3. (a)Following Example 3.1,

(i) find the value of R^2.

(ii) Find the fitted regression equation.

(b) Following Example 3.2,

(i) find the value of R^2.

(ii) Find the fitted regression equation and interpret.

Solution :

(a) (i) From the calculation table of Example 3.1. we get $\Sigma x_{1i}y_i = 20$, $\Sigma x_{2i}y_i = -3$ and $\Sigma y_i^2 = 24$. Further we obtained $\hat{\beta}_0 = 5.368$, $\hat{\beta}_1 = 0.806$ and $\hat{\beta}_2 = -0.720$

We know that $R^2_{Y\cdot X_1,X_2} = R^2 = \dfrac{\text{ESS}}{\text{TSS}} = \dfrac{\hat{\beta}_1\Sigma x_{1i}y_i + \hat{\beta}_2\Sigma x_{2i}y_i}{\Sigma y_i^2}$

$$= \frac{0.806 \times 20 + (-0.72) \times -3}{24} = \frac{16.12 + 2.16}{24} = \frac{18.28}{24} = 0.76$$

$\therefore\ R^2 = 0.76$

(ii) The estimated (fitted) regression equation is

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$$

or, $\hat{Y}_i = 5.368 + 0.806\ X_{1i} - 0.72\ X_{2i}$ and $R^2 = 0.76$

(b) (i) From the calculation table of Example 3.2.

we get $\Sigma x_{1i}y_i = 62$, $\Sigma x_{2i}y_i = 52$, $\Sigma y_i^2 = 272$.

Further we obtained $\hat{\beta}_0 = -23.75$, $\hat{\beta}_1 = -0.25$ and $\hat{\beta}_2 = 5.5$

we thus get $R^2 = \dfrac{\hat{\beta}_1\Sigma x_{1i}y_i + \hat{\beta}_2\Sigma x_{2i}y_i}{\Sigma y_i^2}$

$$= \frac{-0.25 \times 62 + 5.5 \times 52}{272} = \frac{-15.5 + 286}{272} = \frac{270.5}{272} = 0.994$$

$\therefore\ R^2 = 0.994.$

(ii) Thus the estimated regression equation is

$$\hat{Y}_i = -23.75 - 0.25X_{1i} + 5.5X_{2i},\ R^2 = 0.994$$

This equation suggests that years of experience with the firm is far more important than years of education (which actually has a negative sign). This equation says that we can predict that one more year of experience, after allowing for years of education (or, holding it constant), results in an annual increase in salary of \$5500. This means that if we consider the persons with the same level of education, the one with one more year of experience can be expected to have a higher salary of \$5500. Similarly, if we consider two persons with the same experience, the one with an education of one more year can be expected to have a lower annual salary of \$250.

Here $R^2 = 0.995$ implies that out of 100% variation in salary of the employee 99% variation can be explained by the two explanatory variables X_1 and X_2 jointly.

3.7. R^2 and the Adjusted R^2

An important property of R^2 is that it is a non-decreasing function of the number of explanatory variables or regressors present in the model, unless the added variable is perfectly collinear with the other regressors ; as the number of regressors increases, R^2 almost invariably increases and never decreases. Stated differently, an additional X variable will not decrease R^2.

Since $R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\Sigma e_i^2}{\Sigma y_i^2}$

$[\Sigma y_i^2 = \Sigma \hat{y}_i^2 + \Sigma e_i^2 \Rightarrow \text{TSS} = \text{ESS} + \text{RSS}]$

Now, TSS = Σy_i^2 is independent of the number of variables in the model because it is simply $\Sigma(Y_i - \overline{Y})^2$.

The RSS, Σe_i^2, however, depends on the number of regressors present in the model. Intuitively it is clear that as the number of X variales increases, Σe_i^2 is likely to decrease (at least will not increase) ; hence R^2 will increase. In view of this, in comparing two regression models with the same dependent variable but differeing number of X variables, one should be very wary of choosing the model with the highest R^2.

To compare two R^2 terms, one must take into account the number of X variables present in the model. This can be done readily if we consider an alternative coefficient of determination,

$\overline{R}^2 = 1 - \frac{\Sigma e_i^2 / \{n-(K+1)\}}{\Sigma y_i^2 / (n-1)}$ where K = number of explanatory variables and $(K + 1)$ = number of parameters in the model including the intercept term. In a three variable (with two explanatory variables) linear regression model, $K = 2$, so $n - (K + 1) = n - 3$.

The R^2 thus defined is known as adjusted R^2, denoted by $\overline{R}^2$. The term adjusted means adjusted for the degrees of freedom (d.f) associated with the sums of squares of Σe_i^2 and Σy_i^2.

RSS $= \Sigma e_i^2$ has $n - (K + 1)$ degrees of freedom in a model involving $(K + 1)$ paramaters, including the intercept term and TSS $= \Sigma y_i^2$ has $(n - 1)$ degrees of freedom.

Thus the adjusted R^2 can also be written as

$$\overline{R}^2 = 1 - \frac{\hat{\sigma}_u^2}{S_Y^2}$$

where $\hat{\sigma}_u^2 = \Sigma e_i^2 / \{n-(K+1)\}$ is the residual variance, and unbiased estimator of true σ_u^2 and

$$S_Y^2 = \frac{1}{n-1}\Sigma(Y_i - \overline{Y})^2 = \frac{1}{n-1}\Sigma y_i^2 = \text{sample variance of } Y$$

$$\therefore \Sigma y_i^2 = (n-1)S_Y^2 \text{ and } \Sigma y_i^2 / (n-1) = S_Y^2$$

It is thus clear that $\overline{R}^2$ and R^2 are related and we can express the relation as follows :

Since $R^2 = 1 - \frac{\Sigma e_i^2}{\Sigma y_i^2}$ and $\overline{R}^2 = 1 - \frac{\Sigma e_i^2/(n-(K+1))}{\Sigma y_i^2/(n-1)}$

$$\therefore \overline{R}^2 = 1 - \frac{\Sigma e_i^2 \cdot (n-1)}{\Sigma y_i^2[n-(K+1)]} \quad = 1-(1-R^2)\frac{n-1}{n-(K+1)}$$

$$\left[\text{Since } R^2 = 1 - \frac{\Sigma e_i^2}{\Sigma y_i^2} \therefore (1-R^2)\frac{(n-1)}{n-(K+1)} = \frac{\Sigma e_i^2(n-1)}{\Sigma y_i^2[n-(K+1)]}\right]$$

In the case of three variable linear regression model we have $K = 2$ and $(K + 1) = 3$

$$\therefore \overline{R}^2 = 1-(1-R^2)\frac{n-1}{n-3}.$$

From this relation it is clear that for $(K + 1) > 1$, $\overline{R}^2 < R^2$ which implies that as the number of X variables increases, the adjusted R^2 increases less than unadjusted R^2 and $\overline{R}^2$ can be negative, although R^2 is necessarily non negative. In case $\overline{R}^2$ turns out to be negative in an application, its value is taken as zero.

It should be noted that if n is large $\overline{R}^2$ and R^2 will not differ much. But with small samples, if the number of regressors (X's) is large in relation to the sample observations, $\overline{R}^2$ will be much smaller than R^2 and can even assume negative values, in which case $\overline{R}^2$ should be interpreted as being equal to zero.

Note : Comparing Two R^2 values

It is crucial to note that in comparing two models on the basis of the coefficient of determination, whether adjusted or not, the sample size n and the dependent variable must be the same, the explanatory variables may take any form. Thus for the models,

$$\log Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \qquad \text{......... (A)}$$

and $$Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + u_i \qquad \text{......... (B)}$$

the computed R^2 terms cannot be compared. The reason is that by definition, R^2 measures the proportion of the variation in the dependent variable accounted for the explainatory variables. Therefore in equation (A), R^2 measures the proportion of the variation in log Y explained by X_1 and X_2, whereas in equation (B) it measures the proportion of the variation in Y and hence the two are not the same thing. A change in log Y gives a relative or proportional change in Y, whereas a change in Y gives an absolute change. Therefore, $\text{var}(\hat{Y}_i)/\text{var}(Y_i)$ is not equal to $\text{var}(\log \hat{Y}_i)/\text{var}(\log Y_i)$. Thus the two coefficients of determination are not the same.

Example 3.4. (a) Following Example 3.1 find the value of Adjusted R^2.
(b) Following Example 3.2 find the value of Adjusted R^2.

Solution : (a) We know that in a three-variable linear regression model, adjusted R^2, is denoted by

$\overline{R}^2 = 1-(1-R^2)\frac{n-1}{n-3}$. Here we see that following data of Example 3.1, $R^2 = 0.76$, $n = 5$

$$\therefore \bar{R}^2 = 1-(1-0.76)\frac{(5-1)}{(5-3)} = 1-\frac{0.24\times 4}{2} = 1-0.48 = 0.52$$

Thus the value of adjusted $R^2 = \bar{R}^2 = 0.52$

Here the see that $\bar{R}^2 < R^2$ i.e., the value of adjusted R^2 is lower than the unadjusted R^2.

(b) Following Example 3.2 the value of adjusted R^2 is given by $\bar{R}^2 = 1-(1-R^2)\frac{n-1}{n-3}$.

Here $R^2 = 0.995$ and $n = 5$

$$\therefore \bar{R}^2 = 1-(1-R^2)\cdot\frac{n-1}{n-3} = 1-(1-0.994)\times\frac{5-1}{5-3}$$

$= 1-0.012 = 0.988.$

$\therefore$ Adjusted $R^2 = \bar{R}^2 = 0.988$ which is smaller than unadjusted $R^2 = 0.994$.

The value of adjusted $R^2 = \bar{R}^2$ can also be obtained by using the formula

$$\bar{R}^2 = 1-\frac{\Sigma e_i^2/\{n-(K+1)\}}{\Sigma y_i^2/(n-1)}$$

From Example 3.2 we have obtained

$\Sigma e_i^2 = 1.5,\ \Sigma y_i^2 = 272,\ n = 5,\ K = 2,\ K+1 = 3$

$$\therefore \bar{R}^2 = 1-\frac{\frac{1.5}{5-3}}{\frac{272}{5-1}} = 1-\frac{0.75}{68} = 1-0.0110 = 0.988$$

$\therefore \bar{R}^2 = 0.988\cdot$

3.8. Partial Correlation Coefficients and the Coefficient of Partial Determination

In the simple correlation analysis the coefficient of correlation r is used as a measure of the degree of linear association between two variables X and Y

$$[Y_i = \alpha + \beta X_i + u_i,\ i = 1, 2, \ldots, n]$$

For three variable linear regression model $[Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i,\ i = 1, 2, \ldots, n]$ we can compute three correlation coefficients $r_{YX_1} = r_{12}$ (correlation coefficient between Y and X_1), $r_{YX_2} = r_{13}$ (correlation coefficient between Y and X_2) and $r_{X_1X_2} = r_{23}$ (correlation coefficient between X_1 and X_2)

These correlation coefficients are called gross or simple correlation coefficients or correlation coefficients of zero order and computed by the formula

$$r_{XY} = \frac{\text{cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{\frac{1}{n}\Sigma(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\frac{1}{n}\Sigma(X_i-\bar{X})^2}\sqrt{\frac{1}{n}\Sigma(Y_i-\bar{Y})^2}} = \frac{\Sigma x_i y_i}{\sqrt{\Sigma x_i^2}\cdot\sqrt{\Sigma y_i^2}}$$

where $x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$.

But now consider this question ; Does say, r_{YX_1} or r_{12}, in fact, measure the "true" degree of linear association between Y and X_1 when a third variable X_2 may be associated with both of them ?

This question is analogous to the following question : Suppose the true regression model is $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ but we omit from the model the variable X_2 and simply regress Y on X_1, obtaining the slope coefficient of say, b_{12}. Will this coefficient be equal to the true coefficient β_1 ? In general r_{YX_1} or r_{12} is not likely to reflect the true degree of association between Y and X_1 in presence of X_2. As a matter of fact, it is likely to give a false impression of the nature of association between Y and X_1. Therefore what we need is a correlation coefficient that is independent of the influence, if any, of X_2 on X_1. Such a correlation coefficient can be obtained and is known as the **Partial correlation coefficient.**

Conceptually, it is similar to the partial regression coefficient. We define $r_{12.3}$ = Partial correlation coefficient between Y and X_1, holding X_2 constant.

$r_{13.2}$ = Partial correlation coefficient between Y and X_2, holding X_1 constant.

$r_{23.1}$ = Partial correlation coefficient between X_1 and X_2, holding Y constant.

These partial correlations can be obtained from the simple or zero-order correlation coefficients as follows (proofs are not given here) :

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{\left(1-r_{13}^2\right)\left(1-r_{23}^2\right)}},\quad r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{\left(1-r_{12}^2\right)\left(1-r_{23}^2\right)}} \text{ and } r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{\left(1-r_{12}^2\right)\left(1-r_{13}^2\right)}}$$

The partial correlations defined above are called first order correlation coefficients. By order we mean the number of secondary subscripts. Thus $r_{12.34}$ would be the correlation coefficient of order two, $r_{12.345}$ would be the correlation coefficient of order three, and so on. The interpretation of say, $r_{12.34}$ is that it gives the coefficient of correlation between Y and X_1, holding X_2 and X_3 constant.

In the two variable case the simple correlation coefficient r has a straight forward meaning. It measures the degree of (linear) association between the dependent variable Y and the single explanatory variable X. But once we go beyond the two variable case, we need to pay careful attention to the interpretation of the simple correlation coefficient.

From the formula :

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{\left(1-r_{13}^2\right)\left(1-r_{23}^2\right)}}$$, for example, we may observe the following :

(i) Even if $r_{12} = 0$, $r_{12.3}$ will not be zero unless r_{13} or r_{23} or both are zero.

(ii) If $r_{12} = 0$ and r_{13} and r_{23} are non zero and are of the same sign, $r_{12.3}$ will be negative, whereas if they are of the opposite signs, it will be positive.

(iii) The terms $r_{12.3}$ and r_{12} need not have the same sign.

(iv) In the two variable case r^2 lies between 0 and 1 $(0 \leq r^2 \leq 1)$. The same property holds true for partial correlation coefficients,

i.e., $0 \leq r_{12.3}^2 \leq 1$. Similarly, $0 \leq r_{13.2}^2 \leq 1$ and $0 \leq r_{23.1}^2 \leq 1$.

(v) If $r_{13} = r_{23} = 0$, it does not mean that r_{12} is also zero. In other words, the fact that Y and X_2 and X_1 and X_2 are uncorrelated does not mean that Y and X_1 are uncorrelated.

The expression $r^2_{12.3}$ may be called the **coefficient of partial determination** and may be interpreted as the proportion of the variation in Y not explained by the variable X_2 that has been explained by the inclusion of X_1 into the model. Conceptually it is similar to R^2 (multiple coefficient of determination).

The following relations are found between R^2, simple correlation coefficients and partial correlation coefficients :

$$R^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$

$$R^2 = r_{12}^2 + \left(1 - r_{12}^2\right) r_{13.2}^2$$

$$R^2 = r_{13}^2 + \left(1 - r_{13}^2\right) r_{12.3}^2$$

It has ben pointed out earlier that R^2 will not decrease if an additional explanatory variable is introduced into the model, which can be seen clearly from the equation, $R^2 = r_{12}^2 + \left(1 - r_{12}^2\right) r_{13.2}^2$. This equation states that the proportion of the variation in Y explained by X_1 and X_2 jointly is the sum of two parts : the part explained by X_1 alone $\left(= r_{12}^2\right)$ and the part not explained by $X_1 \left(= 1 - r_{12}^2\right)$ times the proportion that is explained by X_2 after holding the influence of X_1 constant. Now $R^2 > r_{12}^2$ so long as $r_{13.2}^2 > 0$. At worst, $r_{13.2}^2$ will be zero, in which case $R^2 = r_{12}^2$.

Example 3.5. Following Example 3.1,

(i) find the values of r_{12}, r_{13} and r_{23}.

(ii) Find the values of the partial regression coefficient $r_{12.3}$.

(iii) Find the value of R^2 in terms of r_{12}, r_{13} and r_{23}.

Solution : From the calculation table of Example 3.1 we have the following values:

$\Sigma x_{1i}^2 = 40$, $\Sigma x_{2i}^2 = 23.20$, $\Sigma y_i^2 = 24$, $\Sigma x_{1i} x_{2i} = 17$, $\Sigma x_{1i} y_i = 20$ and $\Sigma x_{2i} y_i = -3$

where $x_{1i} = X_{1i} - \overline{X}_1$, $x_{2i} = X_{2i} - \overline{X}_2$ and $y_i = Y_i - \overline{Y}$

(i) Now by using the formule of r_{12}, r_{13} and r_{23} and putting the required values we can get the value of r_{12}, r_{13} and r_{23}.

By definition, $r_{12} = r_{YX_1} = \dfrac{\operatorname{cov}(X_1, Y)}{\sigma_{X_1} \cdot \sigma_Y} = \dfrac{\frac{1}{n}\Sigma(X_{1i} - \overline{X}_1)(Y_i - \overline{Y})}{\sqrt{\frac{1}{n}\Sigma(X_{1i} - \overline{X}_1)^2}\sqrt{\frac{1}{n}\Sigma(Y_i - \overline{Y})^2}}$

$$= \frac{\frac{1}{n}\Sigma x_{1i} y_i}{\sqrt{\frac{1}{n}\Sigma x_{1i}^2}\sqrt{\frac{1}{n}\Sigma y_i^2}} = \frac{\Sigma x_{1i} y_i}{\sqrt{\Sigma x_{1i}^2}\sqrt{\Sigma y_i^2}} = \frac{20}{\sqrt{40 \times 24}} = 0.64$$

$\therefore r_{12} = 0.64$

Similarly, $r_{13} = r_{YX_2} = \dfrac{\Sigma x_{2i} y_i}{\sqrt{\Sigma x_{2i}^2}\sqrt{\Sigma y_i^2}} = \dfrac{-3}{\sqrt{23.20 \times 24}} = -0.13$

and $r_{23} = r_{X_1X_2} = \dfrac{\Sigma x_{1i} x_{2i}}{\sqrt{\Sigma x_{1i}^2}\sqrt{\Sigma x_{2i}^2}} = \dfrac{17}{\sqrt{40 \times 23.20}} = 0.56$

$\therefore r_{12} = 0.64,\ r_{13} = -0.13$ and $r_{23} = 0.56$

(ii) We know that $r_{12.3} = \dfrac{r_{12} - r_{13}r_{23}}{\sqrt{\left(1-r_{13}^2\right)\left(1-r_{23}^2\right)}}$

$$= \frac{0.64-(-0.13)\times 0.56}{\sqrt{\{1-(-0.13)^2\}\{1-(0.56)^2\}}} = \frac{0.7128}{0.8246} = 0.86$$

$\therefore r_{12.3} = 0.86$

(iii) We know that,

$$R^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1-r_{23}^2}$$

$$= \frac{(0.64)^2 + (-0.13)^2 - 2\times 0.64 \times (-0.13) \times 0.56}{1-(0.56)^2}$$

$$= \frac{0.4096 + 0.0169 + 0.093184}{0.6864} = \frac{0.5196}{0.6864} \approx 0.76$$

$\therefore R^2 = 0.76.$

Example 3.6. Are the following data consistent ? Give reasons.

(a) $r_{23} = 0.9,\ r_{13} = -0.2,\ r_{12} = 0.8$

(b) $r_{12} = 0.6,\ r_{23} = -0.9,\ r_{31} = -0.5$

Solution : From the above data set we will first calculate the value of

$R^2 = \dfrac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1-r_{23}^2}$ and will verify whether $0 < R^2 < 1$ or not.

(a) Here we see that

$$R^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1-r_{23}^2}$$

$$= \frac{(0.8)^2 + (-0.2)^2 - 2\times 0.8 \times -0.2 \times 0.9}{1-(0.9)^2} = \frac{0.64 + 0.04 + 0.288}{0.19} = \frac{0.968}{0.19} = 5 > 1$$

$\therefore R^2 = 5$ which is not possible as $0 < R^2 < 1$.

Hence the given information are not consistent.

(b) $R^2 = \dfrac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1-r_{23}^2}$, as $r_{31} = r_{13} = -0.5$

$$= \frac{(0.6)^2 + (-0.5)^2 - 2\times 0.6 \times -0.5 \times -0.9}{1-(-0.9)^2}$$

$$= \frac{0.36+0.25-0.54}{0.19} = \frac{0.07}{0.19} = 0.36$$

$\therefore R^2 = 0.36$ which satisfies the condition $0 < R^2 < 1$. Hence the data set may be consistent.

Example 3.7. From the following data estimate the partial regression coefficients, their standard errors, and the adjusted and unadjusted R^2 values :

$\bar{Y} = 367.693, \quad \bar{X}_1 = 402.760, \quad \bar{X}_2 = 8.0$

$\Sigma(Y_i - \bar{Y})^2 = 66042.269, \quad \Sigma(X_{1i} - \bar{X}_1)^2 = 84855.096$

$\Sigma(X_{2i} - \bar{X}_2)^2 = 280.000, \quad \Sigma(Y_i - \bar{Y})(X_{1i} - \bar{X}_1) = 74778.346$

$\Sigma(Y_i - \bar{Y})(X_{2i} - \bar{X}_2) = 4250.900, \quad \Sigma(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2) = 4796.00$ and $n = 15$

Solution : The given results are related to a three variable linear regression model of the form $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, i = 1, 2, ..., n$

where β_1 and β_2 are the partial regression coefficients. Assuming $Y_i - \bar{Y} = y_i$, $X_{1i} - \bar{X}_1 = x_{1i}$ and $X_{2i} - \bar{X}_2 = x_{2i}$ we can obtain the OLS estimators of β_1 and β_2 as follows :

$$\hat{\beta}_1 = \frac{\Sigma x_{2i}^2 \cdot \Sigma x_{1i} y_i - \Sigma x_{1i} x_{2i} \cdot \Sigma x_{2i} y_i}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2}$$

$$= \frac{280.000 \times 74778.346 - 4796.00 \times 4250.900}{84855.096 \times 280.00 - (4796.00)^2}$$

$$= \frac{20937936.88 - 20387316.40}{23759426.88 - 23001616} = \frac{550620.48}{757810.88} = 0.7265$$

$\therefore \hat{\beta} = 0.7265$

$$\text{Similarly, } \hat{\beta}_2 = \frac{\Sigma x_{1i}^2 \cdot \Sigma x_{2i} y_i - \Sigma x_{1i} x_{2i} \cdot \Sigma x_{1i} y_i}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2}$$

$$= \frac{84855.096 \times 4250.900 - 4796.00 \times 74778.346}{84855.096 \times 280.000 - (4796.00)^2}$$

$$= \frac{2073580.17}{757810.88} = 2.7362$$

$\therefore \hat{\beta}_2 = 2.7362$

Now we have to find out $SE(\hat{\beta}_1) = \sqrt{\text{var}(\hat{\beta}_1)}$ and $SE(\hat{\beta}_2) = \sqrt{\text{var}(\hat{\beta}_2)}$

$$\text{We know that } \text{var}(\hat{\beta}_1) = \frac{\sigma_u^2 \Sigma x_{2i}^2}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2}$$

$$\text{and } \text{var}(\hat{\beta}_2) = \frac{\sigma_u^2 \Sigma x_{1i}^2}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2}$$

Here σ_u^2 is unknown and hence it is to be replaced by its unbiased estimator

$$\hat{\sigma}_u^2 = \frac{\Sigma e_i^2}{n-3}.$$

Since $\Sigma y_i^2 = \Sigma \hat{y}_i^2 + \Sigma e_i^2$ i.e. TSS = ESS + RSS

$\therefore \Sigma e_i^2 = \Sigma y_i^2 - \Sigma \hat{y}_i^2$

where $\Sigma \hat{y}_i^2 = \text{ESS} = \hat{\beta}_1 \Sigma x_{1i} y_i + \hat{\beta}_2 \Sigma x_{2i} y_i$.

Since $\hat{\beta}_1 = 0.7265$, $\hat{\beta}_2 = 2.7362$ and $\Sigma x_{1i} y_i = 74778.346$ and $\Sigma x_{2i} y_i = 4250.90$.

$$\therefore \text{ESS} = \Sigma \hat{y}_i^2 = \hat{\beta}_1 \Sigma x_{1i} y_i + \hat{\beta}_2 \Sigma x_{2i} y_i$$

$$= 0.7265 \times 74778.346 + 2.7362 \times 4250.90$$

$$= 65957.7809$$

Again TSS = $\Sigma y_i^2 = 66042.269$

$$\therefore \Sigma e_i^2 = \Sigma y_i^2 - \Sigma \hat{y}_i^2 = 66042.269 - 65957.7809 = 84.4881$$

$$\therefore \hat{\sigma}_u^2 = \frac{\Sigma e_i^2}{n-3} = \frac{84.4881}{15-3} = \frac{84.4881}{12} = 7.0406$$

Now $\text{var}(\hat{\beta}_1) = \dfrac{\hat{\sigma}_u^2 \Sigma x_{2i}^2}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2}$

$$= \frac{7.0406 \times 280.000}{84855.096 \times 280.000 - (4796.00)^2}$$

$$= \frac{1971.368}{757810.88} = 0.00260$$

$\therefore SE(\hat{\beta}_1) = \sqrt{\text{var}(\hat{\beta}_1)} = 0.0510$

Similarly, $\text{var}(\hat{\beta}_2) = \dfrac{\hat{\sigma}_u^2 \Sigma x_{1i}^2}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2}$

$$= \frac{7.0406 \times 84855.096}{757810.88} = \frac{597430.7889}{757810.88} = 0.78836$$

$\therefore SE(\hat{\beta}_2) = \sqrt{\text{var}(\hat{\beta}_2)} = 0.8878$

Now we have to find out the value of R^2 and adjusted $R^2 = \bar{R}^2$.

We know that $R^2 = \dfrac{\text{ESS}}{\text{TSS}} = 1 - \dfrac{\text{RSS}}{\text{TSS}} = 1 - \dfrac{\Sigma e_i^2}{\Sigma y_i^2}$

[since $\Sigma y_i^2 = \Sigma \hat{y}_i^2 + \Sigma e_i^2 \Rightarrow$ TSS = ESS + RSS]

$$\therefore R^2 = \frac{ESS}{TSS} = \frac{\Sigma\hat{y}_i^2}{\Sigma y_i^2} = \frac{65957.7809}{66042.269} = 0.998$$

$\therefore R^2 = 0.998$

Again, adjusted $R^2 = \bar{R}^2 = 1-(1-R^2)\frac{n-1}{n-3}$. Here $n = 15$

$$\therefore \bar{R}^2 = 1-(1-0.998)\frac{15-1}{15-3} = 0.997$$

So, R^2 and $\bar{R}^2$ are almost the same.

3.9. Confidence Intervals and Hypothesis Testing in a Three Variable Multiple Linear Regression Model

We consider a multiple linear regression model with two independent variables, given by,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i,\ i = 1, 2, ..., n$$

Suppose $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ and $\hat{\sigma}_u^2$ are the OLS estimators of $\beta_0, \beta_1, \beta_2$ and σ_u^2 respectively.

We also know that

$\hat{\beta}_1 \sim N[\beta_1, \text{var}(\hat{\beta}_1)]$ where $E(\hat{\beta}_1) = \beta_1$ and $\text{var}(\hat{\beta}_1) = \dfrac{\sigma_u^2}{n\sigma_{X_1}^2\left[1-r_{X_1X_2}^2\right]}$. This is read as $\hat{\beta}_1$ is normally distributed with mean β_1 and variance var $(\hat{\beta}_1)$.

Similarly, $\hat{\beta}_2 \sim N[\beta_2, \text{var}(\hat{\beta}_2)]$ where $E(\hat{\beta}_2) = \beta_2$

$$\text{and } \text{var}(\hat{\beta}_2) = \frac{\sigma_u^2}{n\sigma_{X_2}^2\left[1-r_{X_1X_2}^2\right]}$$

$\hat{\beta}_0 \sim N\left[\beta_0, \text{var}(\hat{\beta}_0)\right]$ where $E(\hat{\beta}_0) = \beta_0$

$$\text{and } \text{var}(\hat{\beta}_0) = \frac{\sigma_u^2}{n} + \bar{X}_1^2\,\text{var}(\hat{\beta}_1) + 2\bar{X}_1\bar{X}_2\,\text{cov}(\hat{\beta}_1, \hat{\beta}_2) + \bar{X}_2^2\,\text{var}(\hat{\beta}_2)$$

and $u \sim N(0, \sigma_u^2)$ where $E(u) = 0$ and var $(u) = \sigma_u^2$

$$\text{and } \text{cov}(\hat{\beta}_1, \hat{\beta}_2) = \frac{-\sigma_u^2 r_{X_1X_2}}{n\sigma_{X_1}\sigma_{X_2}(1-r_{X_1X_2}^2)}$$

We are now interested in testing the following hypothesis :

Case 1 : We want to test the null hypothesis $H_0 : \beta_0 = 0$ against the alternative hypothesis, either $H_1 : \beta_0 \neq 0$ or $H_1 : \beta_0 > 0$ or $H_1 : \beta_0 < 0$

Since, $\hat{\beta}_0 \sim N[\beta_0, \text{var}(\hat{\beta}_0)]$

$$\text{Now } \tau \text{ or } Z = \frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)} \sim N(0,1)$$

would be the appropriate test statistic where $SE(\hat{\beta}_0) = \sqrt{\text{var}(\hat{\beta}_0)}$.

When σ_u^2 is unknown and it is replaced by its unbiased estimator $\hat{\sigma}_u^2 = \Sigma e_i^2/_{n-3}$

$\left[\text{Since } E\left[\frac{\Sigma e_i^2}{n-3}\right] = \sigma_u^2\right]$ then the appropriate test statistic would be $t = \frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)}$ which will follow a t distribution with d.f. = $(n - 3)$. Under $H_0 : \beta_0 = 0$, the test statistic would be

$$t = \frac{\hat{\beta}_0 - 0}{SE(\hat{\beta}_0)} = \frac{\hat{\beta}_0}{SE(\hat{\beta}_0)} \sim t_{n-3}$$

Nature of the Test :

(i) For the alternative hypothesis $H_1 : \beta_0 \neq 0$, the null hypothesis $H_1 : \beta_0 = 0$ will be accepted at 100α% level of significance if for the given sample $-t_{\alpha/2,n-3} \leq t \leq t_{\alpha/2,n-3}$ and will be rejected otherwise. [i.e. when $|t| > t_{\alpha/2,n-1}$].

(ii) For the alternative hypothesis $H_1 : \beta_0 > 0$, the null hypothesis $H_1 : \beta_0 = 0$ will be accepted if for the given sample $t \leq t_{\alpha,n-3}$ and will be rejected otherwise. [i.e. when $t > t_{\alpha,n-3}$].

(iii) For the althernative hypothesis $H_1 : \beta_0 < 0$, the null hypothesis $H_1 : \beta_0 = 0$ will be accepted if for the given sample $-t_{\alpha,n-3} \leq t$ and will be rejected otherwise. [i.e. when $t < -t_{\alpha,n-3}$]. In each case α denotes the chosen level of significance. Usually, α = 0.01 or 0.05.

Confidence interval for β_0:

As regards the problem of interval estimation of β_0 at 100 α% level of significance the confidence limits to β_0 would be given by

$$\hat{\beta}_0 \pm t_{\alpha/2,n-3} SE(\hat{\beta}_0)$$

i.e., $P[-t_{\alpha/2,n-3} \leq t \leq t_{\alpha/2,n-3}] = 1 - \alpha$

or, $P\left[-t_{\alpha/2,n-3} \leq \frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)} \leq t_{\alpha/2,n-3}\right] = 1 - \alpha$

or, $P\left[\hat{\beta}_0 - t_{\alpha/2,n-3} SE(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2,n-3} SE(\hat{\beta}_2)\right] = 1 - \alpha$

Here (1 − α) is called the confidence coefficient.

Case 2 : We want to test the null hypothesis $H_0 : \beta_1 = 0$ against the alternative hypothesis $H_1 : \beta_1 \neq 0$ or, $H_1 : \beta_1 > 0$ or $H_1 : \beta_1 < 0$.

Since $\hat{\beta}_1 \sim N[\beta_1, \text{var}(\hat{\beta}_1)]$ where $E(\hat{\beta}_1) = \beta_1$

and $\text{var}(\hat{\beta}_1) = \frac{\sigma_u^2}{n\sigma_{X_1}^2 (1 - r_{X_1X_2}^2)}$

Now τ or $Z = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim N(0,1)$ would be the appropriate test statistic where

$SE(\hat{\beta}_1)=\sqrt{\text{var}(\hat{\beta}_1)}$. When σ_u^2 is not known then it is replaced by its unbiased estimator, $\hat{\sigma}_u^2=\Sigma e_i^2/(n-3)$. Then the test statistic becomes $t=\dfrac{\hat{\beta}_1-\beta_1}{SE(\hat{\beta}_1)}\sim t_{n-3}$.

Under $H_0:\beta_1=0$, the test statistic would be

$$t=\frac{\hat{\beta}_1-0}{SE(\hat{\beta}_1)}=\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}\sim t_{n-3}$$

Nature of the Test :

(i) For the alternative hypothesis $H_1:\beta_1\neq 0$, the null hypothesis $H_0:\beta_1=0$ will be accepted if for the given sample $-t_{\alpha/2,n-3}\leq t\leq t_{\alpha/2,n-3}$ and will be rejected otherwise.

(ii) For the alternative hypothesis $H_1:\beta_1>0$, the null hypothesis $H_0:\beta_1=0$ will be accepted if for the given sample $t\leq t_{\alpha,n-3}$ and will be rejected otherwise. [i.e. when $t>t_{\alpha,n-3}$].

(iii) For the alternative hypothesis $H_1:\beta_1<0$, the null hypothesis $H_0:\beta_1=0$ will be accepted if for the given sample $-t_{\alpha,n-3}\leq t$ and will be rejected otherwise. [i.e. when $t>-t_{\alpha,n-3}$]. In each case α (= 0.01 or 0.05) denotes the chosen level of significance.

Confidence Interval for β_1:

As regards the problem of interval estimation of β_1 at 100 α% level of significance, the confidence limits to $\beta_1$would be given by,

$$\hat{\beta}_1\pm t_{\alpha/2,n-3}SE(\hat{\beta}_1)$$

i.e., $$P[-t_{\alpha/2,n-3}\leq t\leq t_{\alpha/2,n-3}]=1-\alpha$$

or, $$P\left[-t_{\alpha/2,n-3}\leq\frac{\hat{\beta}_1-\beta_1}{SE(\hat{\beta}_1)}\leq t_{\alpha/2,n-3}\right]=1-\alpha$$

or, $$P\left[\hat{\beta}_1-t_{\alpha/2,n-3}SE(\hat{\beta}_1)\leq\beta_1\leq\hat{\beta}_1+t_{\alpha/2,n-3}SE(\hat{\beta}_1)\right]=1-\alpha$$

Here $(1-\alpha)$ is the confidence coefficient.

Case 3 : We want to test the null hypothesis $H_0:\beta_2=0$ against the alternative hypothesis, $H_1:\beta_2\neq 0$ or $H_1:\beta_2>0$ or, $H_1:\beta_2<0$

Since $\hat{\beta}_2\sim N[\beta_2,\text{var}(\hat{\beta}_2)]$ where $E(\hat{\beta}_2)=\beta_2$

$$\text{and}\quad \text{var}(\hat{\beta}_2)=\frac{\sigma_u^2}{n\sigma_{X_2}^2(1-r_{X_1X_2}^2)}$$

Now τ or $Z=\dfrac{\hat{\beta}_2-\beta_2}{SE(\hat{\beta}_2)}\sim N(0,1)$ would be the appropriate test statistic where $SE(\hat{\beta}_2)=\sqrt{\text{var}(\hat{\beta}_2)}$. When σ_u^2 is not known then it is replaced by its unbiased estimator $\hat{\sigma}_u^2=\Sigma e_i^2/(n-3)$ and the test statistic becomes $t=\dfrac{\hat{\beta}_2-\beta_2}{SE(\hat{\beta}_2)}\sim t_{n-3}$.

Under $H_1 : \beta_2 = 0$, the test statistic would be,

$$t = \frac{\hat{\beta}_2 - 0}{SE(\hat{\beta}_2)} = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} \sim t_{n-3}$$

Nature of the Test :

(i) For the alternative hypothesis $H_1 : \beta_2 \neq 0$, the null hypothesis $H_0 : \beta_2 = 0$ will be accepted if for the given sample $-t_{\alpha/2, n-3} \leq t \leq t_{\alpha/2, n-3}$ and will be rejected otherwise.

(ii) For the alternative hypothesis $H_1 : \beta_2 > 0$, the null hypothesis $H_0 : \beta_1 = 0$ will be accepted if for the given sample $t \leq t_{\alpha, n-3}$ and will be rejected otherwise. [i.e. when $t > t_{\alpha, n-3}$]

(iii) For the alternative hypothesis $H_1 : \beta_2 < 0$, the null hypothesis $H_0 : \beta_2 = 0$ will be accepted if for the given sample $-t_{\alpha, n-3} \leq t$ and will be rejected otherwise [i.e. when $t < -t_{\alpha, n-3}$]

Confidence interval for β_2:

As regards the problem of interval estimation of β_2 at 100 α% level of significance, the confidence limits to β_2 would be given by,

$$\hat{\beta}_2 \pm t_{\alpha/2, n-3} SE(\hat{\beta}_2)$$

i.e., $P[-t_{\alpha/2, n-3} \leq t \leq t_{\alpha/2, n-3}] = 1 - \alpha$

or, $$P\left[-t_{\alpha/2, n-3} \leq \frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} \leq t_{\alpha/2, n-3}\right] = 1 - \alpha$$

or, $$P\left[\hat{\beta}_2 - t_{\alpha/2, n-3} SE(\hat{\beta}_2) \leq \beta_2 \leq \hat{\beta}_2 + t_{\alpha/2, n-3} SE(\hat{\beta}_2)\right] = 1 - \alpha$$

where $(1 - \alpha)$ is the confidence coefficient.

Case 4 : We want to test the null hypothesis $H_0 : \beta_1 = \beta_2$ against the alternative hypothesis, $H_1 : \beta_1 \neq \beta_2$ or $H_1 : \beta_1 > \beta_2$ or, $H_1 : \beta_1 < \beta_2$

Since $(\hat{\beta}_1 - \hat{\beta}_2) \sim N[(\beta_1 - \beta_2), \text{var}(\hat{\beta}_1 - \hat{\beta}_2)]$

where $E(\hat{\beta}_1 - \hat{\beta}_2) = E(\hat{\beta}_1) - E(\hat{\beta}_2) = \beta_1 - \beta_2$

and $\text{var}(\hat{\beta}_1 - \hat{\beta}_2) = \text{var}(\hat{\beta}_1) + \text{var}(\hat{\beta}_2) - 2\,\text{cov}(\hat{\beta}_1, \hat{\beta}_2)$

$$= \frac{\sigma_u^2}{n\sigma_{X_1}^2 (1 - r_{X_1X_2}^2)} + \frac{\sigma_u^2}{n\sigma_{X_2}^2 (1 - r_{X_1X_2}^2)} + 2\frac{\sigma_u^2 r_{X_1X_2}}{n\sigma_{X_1}\sigma_{X_2}(1 - r_{X_1X_2}^2)}$$

as $$Cov(\hat{\beta}_1, \hat{\beta}_2) = \frac{-\sigma_u^2 r_{X_1X_2}}{n\sigma_{X_1} n\sigma_{X_2}(1 - r_{X_1X_2}^2)} = \frac{\sigma_u^2}{n\left[1 - r_{X_1X_2}^2\right]}\left[\frac{1}{\sigma_{X_1}^2} + \frac{1}{\sigma_{X_2}^2} + \frac{2r_{X_1X_2}}{\sigma_{X_1}\sigma_{X_2}}\right]$$

The appropriate test statistic would be given by,

$$\tau \text{ or } Z = \frac{(\hat{\beta}_1 - \hat{\beta}_2) - (\beta_1 - \beta_2)}{SE(\hat{\beta}_1 - \hat{\beta}_2)} \sim N(0, 1)$$

where $SE(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{\text{var}(\hat{\beta}_1 - \hat{\beta}_2)}$.

If σ_u^2 is unknown and it is replaced by its unbiased estimator $\hat{\sigma}_u^2 = \Sigma e_i^2/(n-3)$,

then the test statistic would be $t = \dfrac{(\hat{\beta}_1 - \hat{\beta}_2) - (\beta_1 - \beta_2)}{SE(\hat{\beta}_1 - \hat{\beta}_2)} \sim t_{n-3}$.

In calculating $SE(\hat{\beta}_1 - \hat{\beta}_2)$ we will just replace $\Sigma e_i^2/(n-3)$ in place of σ_u^2 and other things will remain the same.

The test statistic under $H_0 : \beta_1 - \beta_2 = 0$

or, $H_0 : \beta_1 = \beta_2$ would be $t = \dfrac{\hat{\beta}_1 - \hat{\beta}_2}{SE(\hat{\beta}_1 - \hat{\beta}_2)} \sim t_{n-3}$.

Nature of the Test :

(i) For the alternative hypothesis $H_1 : \beta_1 \neq \beta_2$, the null hypothesis $H_0 : \beta_1 = \beta_2$ will be accepted if for the given sample $-t_{\alpha/2, n-3} \leq t \leq t_{\alpha/2, n-3}$ and will be rejected otherwise.

(ii) For the alternative hypothesis $H_1 : \beta_1 > \beta_2$, the null hypothesis $H_0 : \beta_1 = \beta_2$ will be accepted if for the given sample $t \leq t_{\alpha, n-3}$ and will be rejected otherwise. [i.e. when $t > t_{\alpha, n-3}$].

(iii) For the alternative hypothesis $H_1 : \beta_1 < \beta_2$, the null hypothesis $H_0 : \beta_1 = \beta_2$ will be accepted if for the given sample $-t_{\alpha, n-3} \leq t$ and will be rejected otherwise. [i.e. when $t < -t_{\alpha, n-3}$]. In each use α denotes the chosen level of significance.

Confidence interval of $(\beta_1 - \beta_2)$:

At 100 α% level of significance, the confidence limits to $(\beta_1 - \beta_2)$ would be given by,

$$\left(\hat{\beta}_1 - \hat{\beta}_2\right) \pm t_{\alpha/2, n-3} SE\left(\hat{\beta}_1 - \hat{\beta}_2\right)$$

i.e., $P[-t_{\alpha/2, n-3} \leq t \leq t_{\alpha/2, n-3}] = 1 - \alpha$

$$\text{or,} \quad P\left[-t_{\alpha/2, n-3} \leq \frac{(\hat{\beta}_1 - \hat{\beta}_2) - (\beta_1 - \beta_2)}{SE(\hat{\beta}_1 - \hat{\beta}_2)} \leq t_{\alpha/2, n-3}\right] = 1 - \alpha$$

$$\text{or,} \quad P\left[(\hat{\beta}_1 - \hat{\beta}_2) - t_{\alpha/2, n-3} SE(\hat{\beta}_1 - \hat{\beta}_2) \leq (\beta_1 - \beta_2) \leq (\hat{\beta}_1 - \hat{\beta}_2) + t_{\alpha/2, n-3} SE(\hat{\beta}_1 - \hat{\beta}_2)\right]$$

$= 1 - \alpha$

where $(1 - \alpha)$ is the confidence coefficient.

Case 5 : Confidence interval for σ_u^2:

Under the normality assumption, the variable

$$\chi^2 = \frac{RSS}{\sigma_u^2} = \frac{\Sigma e_i^2}{\sigma_u^2} = (n-3)\frac{\hat{\sigma}_u^2}{\sigma_u^2}$$

follows a χ^2 (chi-square) distribution with d.f = n −3 [where $\hat{\sigma}_u^2 = \Sigma e_i^2/(n-3)$, is an unbiased estimator of σ_u^2]

Therefore we can use χ^2_{n-3} to establish a confidence interval for σ^2_u.

At 100 α% level of significance the confidence limits to σ^2_u would be given by,

$(n-3)\dfrac{\hat{\sigma}^2_u}{\chi^2_{\alpha/2}}$ and $(n-3)\dfrac{\hat{\sigma}^2_u}{\chi^2_{1-\alpha/2}}$ where χ^2 values are taken from the table with d.f. $= (n-3)$.

i.e., $$P\left[\chi^2_{1-\alpha/2} \le \chi^2_{n-3} \le \chi^2_{\alpha/2}\right] = 1-\alpha$$

or, $$P\left[\chi^2_{1-\alpha/2} \le (n-3)\frac{\hat{\sigma}^2_u}{\sigma^2_u} \le \chi^2_{\alpha/2}\right] = 1-\alpha$$

or, $$P\left[(n-3)\frac{\hat{\sigma}^2_u}{\chi^2_{\alpha/2}} \le \sigma^2_u \le (n-3)\frac{\hat{\sigma}^2_u}{\chi^2_{1-\alpha/2}}\right] = 1-\alpha$$

where (1 – α) is the confidence coefficient.

Example 3.8. The following table contains observations on the quantity demanded of a certain commodity (Y), its price (X_1 in $) and consumer's income (X_2 in $)

Y :	100	75	80	70	50	65	90	100	110	60
X_1 :	5	7	6	6	8	7	5	4	3	9
X_2 :	1000	600	1200	500	300	400	1300	1100	1300	300

Assume a linear regression equation of the form

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i,\ i = 1, 2, 3, ..., 10.$$

(i) Find the OLS estimators of β_0, β_1 and β_2 i.e., $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$.

(ii) Find R^2 and adjusted $R^2(\bar{R}^2)$.

(iii) Find $\text{var}(\hat{\beta}_0)$, $\text{var}(\hat{\beta}_1)$ and $\text{var}(\hat{\beta}_2)$.

(iv) Find $SE(\hat{\beta}_0)$, $SE(\hat{\beta}_1)$ and $SE(\hat{\beta}_1)$.

(v) Write the regression results in the summary form.

(vi) Test $H_0 : \beta_0 = 0$ against $H_1 : \beta_0 \neq 0$ and find 95% and 99% confidence intervals for β_0.

(vii) Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ and find 95% and 99% confidence intervals for β_1.

(viii) Test $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$ and find 95% and 99% confidence intervals for β_2.

(ix) Test $H_0 : \beta_1 = \beta_2$ against $H_1 : \beta_1 \neq \beta_2$ and find 95% and 99% confidence intervals for $(\beta_1 - \beta_2)$.

(x) Construct 95% and 99% confidence intervals of σ^2_u.

Calculation Table 3.3

n	Y	X_1	X_2	$y_i = Y_i - \overline{Y}$	$x_{1i} = X_{1i} - \overline{X}_1$	$x_{2i} = X_{2i} - \overline{X}_2$	y_i^2	x_{1i}^2	x_{2i}^2	$x_{1i}y_i$	$x_{2i}y_i$	$x_{1i}x_{2i}$
1	100	5	1000	20	–1	200	400	1	40,000	–20	4000	–200
2	75	7	600	–5	1	–200	25	1	40,000	–5	1000	–200
3	80	6	1200	0	0	400	0	0	160,000	0	0	0
4	70	6	500	–10	0	–300	100	0	90,000	0	3000	0
5	50	8	300	–30	2	–500	900	4	250,000	–60	15000	–1000
6	65	7	400	–15	1	–400	225	1	160,000	–15	6000	–400
7	90	5	1300	10	–1	500	100	1	250,000	–10	5000	–500
8	100	4	1100	20	–2	300	400	4	90,000	–40	6000	–600
9	110	3	1300	30	–3	500	900	9	250,000	–90	15000	–1500
10	60	9	300	–20	3	–500	400	9	250,000	–60	10,000	–1500
n=10	ΣY_i = 800	ΣX_{1i} = 60	ΣX_{2i} = 8000	Σy_i = 0	Σx_{1i} = 0	Σx_{2i} = 0	Σy_i^2 = 3450	Σx_{1i}^2 = 30	Σx_{2i}^2 = 1580,000	$\Sigma x_{1i}.y_i$ =–300	$\Sigma x_{2i}.y_i$ =65000	$\Sigma x_{1i}.x_{2i}$ =–5900

$$\therefore\ \overline{Y} = \frac{\Sigma Y_i}{n} = \frac{800}{10} = 80,\quad \overline{X}_1 = \Sigma X_{1i}/n = \frac{60}{10} = 6 \text{ and } \overline{X}_2 = \frac{\Sigma X_{2i}}{n} = \frac{8000}{10} = 800$$

Solution :

(i) The OLS estimators are as follows :

$$\hat{\beta}_1 = \frac{\Sigma x_{2i}^2 \cdot \Sigma x_{1i} y_i - \Sigma x_{1i} x_{2i} \cdot \Sigma x_{2i} y_i}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2}$$

$$= \frac{1580,000 \times (-300) - (-5900) \times 65000}{30 \times 1580,000 - (-5900)^2} = \frac{-90,50,0000}{125,90000} = -7.19$$

$\therefore \hat{\beta}_1 = -7.19$

Similarly, $\hat{\beta}_2 = \dfrac{\Sigma x_{1i}^2 \cdot \Sigma x_{2i} y_i - \Sigma x_{1i} x_{2i} \cdot \Sigma x_{1i} y_i}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2}$

$$= \frac{30 \times 65,000 - (-5900) \times -300}{30 \times 1580,000 - (-5900)^2} = \frac{180,000}{12590,000} = 0.0143$$

$\therefore \hat{\beta}_2 = 0.0143$

When $\hat{\beta}_1$ and $\hat{\beta}_2$ are known $\hat{\beta}_0$ can be obtained from the relation

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \hat{\beta}_2 \bar{X}_2$$

or, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$ where $\bar{Y} = 80$, $\bar{X}_1 = 6$, $\bar{X}_2 = 800$

$= 80 - (-7.19) \times 6 - 0.0143 \times 800$

$= 80 + 43.14 - 11.44 = 111.70 \quad \therefore \hat{\beta}_0 = 111.70.$

Thus we have $\hat{\beta}_0 = 111.70$, $\hat{\beta}_1 = -7.19$ and $\hat{\beta}_2 = 0.0143$

(ii) We know that $R^2 = \dfrac{\text{ESS}}{\text{TSS}} = \dfrac{\Sigma \hat{y}_i^2}{\Sigma y_i^2} = \dfrac{\hat{\beta}_1 \Sigma x_{1i} y_i + \hat{\beta}_2 \Sigma x_{2i} y_i}{\Sigma y_i^2}$

$$= \frac{-7.19 \times -300 + 0.0143 \times 65000}{3450} = \frac{3086.5}{3450} = 0.894$$

$\therefore R^2 = 0.894$. This means that price and income can jointly explain.89.4% variation in demand out of total variation of 100%.

Now, adjusted $R^2 = \bar{R}^2 = 1 - (1 - R^2) \cdot \dfrac{n-1}{n-3}$, Here $n = 10$.

$\therefore \bar{R}^2 = 1 - (1 - 0.894) \times \dfrac{10-1}{10-3} = 1 - 0.106 \times \dfrac{9}{7} = 1 - 0.1362 = 0.8637$

$\therefore R^2 = 0.894$ and adjusted $R^2 = \bar{R}^2 = 0.8637$

(iii) We know that $\text{var}(\hat{\beta}_1) = \dfrac{\hat{\sigma}_u^2 \Sigma x_{2i}^2}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2}$

Here σ_u^2 is unknown and it is replaced by its unbiased estimator $\hat{\sigma}_u^2 = \dfrac{\Sigma e_i^2}{n-3}$

We know that $R^2 = 1 - \frac{\Sigma e_i^2}{\Sigma y_i^2} = 1 - \frac{\text{RSS}}{\text{TSS}}$

$\therefore \Sigma e_i^2 = \Sigma y_i^2 (1 - R^2) = 3450 \times (1 - 0.894) = 3450 \times 0.106 = 365.7.$

Now $\hat{\sigma}_u^2 = \frac{\Sigma e_i^2}{n-3} = \frac{365.7}{10-3} = \frac{365.7}{7} = 52.24$

Now $\text{var}(\hat{\beta}_1) = \frac{\hat{\sigma}_u^2 \cdot \Sigma x_{2i}^2}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2}$

$$= \frac{52.24 \times 1580,000}{30 \times 1580,000 - (-5900)^2} = \frac{82539200}{12590000} \approx 6.55$$

$\therefore \text{var}(\hat{\beta}_1) = 6.55$

Again, $\text{var}(\hat{\beta}_2) = \frac{\hat{\sigma}_u^2 \cdot \Sigma x_{1i}^2}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2}$

$$= \frac{52.24 \times 30}{30 \times 1580,000 - (-5900)^2} = \frac{1567.20}{12590000} = 0.000124$$

$\therefore \text{var}(\hat{\beta}_2) = 0.000124$

Again, $\text{var}(\hat{\beta}_0) = \frac{\hat{\sigma}_u^2}{n} + \overline{X}_1^2 \cdot \text{var}(\hat{\beta}_1) + 2\overline{X}_1 \overline{X}_2 \text{cov}(\hat{\beta}_1, \hat{\beta}_2) + \overline{X}_2^2 \cdot \text{var}(\hat{\beta}_2)$.

We know that $\hat{\sigma}_u^2 = 52.24$, $n = 10$, $\overline{X}_1 = 6$, $\overline{X}_2 = 800$, $\text{var}(\hat{\beta}_1) = 6.55$, $\text{var}(\hat{\beta}_2) = 0.000124$

and $\text{cov}(\hat{\beta}_1, \hat{\beta}_2) = \frac{-\hat{\sigma}_u^2 \Sigma x_{1i} x_{2i}}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2}$

$$= \frac{-52.24 \times (-5900)}{30 \times 1580,000 - (-5900)^2} = \frac{308216}{12590000} = 0.0245$$

$\therefore \text{cov}(\hat{\beta}_1, \hat{\beta}_2) = 0.0245$. We now put these values in the expression of $\hat{\beta}_0$ and we get,

$$\text{var}(\hat{\beta}_0) = \frac{52.24}{10} + (6)^2 \times 6.55 + 2 \times 6 \times 800 \times 0.0245 + (800)^2 \times 0.000124$$

$$= 5.224 + 235.8 + 235.2 + 79.36 = 555.58$$

(iv) We know that $SE(\hat{\beta}_1) = \sqrt{\text{var}(\hat{\beta}_1)}$

$\therefore SE(\hat{\beta}_1) = \sqrt{\text{var}(\hat{\beta}_1)} = \sqrt{6.55} = 2.5592$.

Similarly, $SE(\hat{\beta}_2) = \sqrt{\text{var}(\hat{\beta}_2)} = \sqrt{0.000124} = 0.0111$

and $SE(\hat{\beta}_0) = \sqrt{\text{var}(\hat{\beta}_0)} = \sqrt{555.58} = 23.570$

(v) The regression results in summary form :

$\hat{Y}_i = 111.70 - 7.19X_{1i} + 0.0143X_{2i}$

$SE : (23.570) \quad (2.5592) \quad (0.0111)$

$R^2 = 0.894$, Adjusted $R^2 = \bar{R}^2 = 0.8637$

(vi) We have to test the null hypothesis $H_0 : \beta_0 = 0$ against the alternative $H_1 : \beta_0 \neq 0$. The appropriate test statistic under $H_0 : \beta_0 = 0$ would be

$$t = \frac{\hat{\beta}_0}{SE(\hat{\beta}_0)} \sim t_{n-3}.$$

Here t (observed) $= \dfrac{\hat{\beta}_0}{SE(\hat{\beta}_0)} = \dfrac{111.70}{23.570} = 4.739$

The null hypothesis $H_0 : \beta_0 = 0$ will be accepted if for the given sample $-t_{\alpha/2,n-3} \leq t \leq t_{\alpha/2,n-3}$ and will be rejected otherwise.

When $\alpha = 0.05$, $t_{\alpha/2,n-3} = t_{0.025,(10-3)} = t_{0.025,7} = 2.365$

$\therefore t_{0.025,7} = 2.365$ (Table value)

Thus we see that t (observed) $= 4.739$ does not lie in the interval $-2.365(-t_{\alpha/2,n-3})$ and $2.365(t_{\alpha/2,n-3})$, Hence $H_0 : \beta_0 = 0$ is rejected and $H_1 : \beta_0 \neq 0$ is accepted at 5% level of significance.

Similarly, when $\alpha = 0.01$, $t_{\alpha/2,n-3} = t_{0.005,7} = 3.499$. Here we see that t (observed) $= 4.739$ does not lie in the interval -3.499 and 3.499. Hence $H_0 : \beta_0 = 0$ is rejected and $H_0 : \beta_0 \neq 0$ is accepted at 1% level of significance.

We know that 100 (1 – α)% confidence interval of β_0 would be :

$$P\left[-t_{\alpha/2,n-3} \leq t \leq t_{\alpha/2,n-3}\right] = 1-\alpha$$

$$\text{or, } P\left[-t_{\alpha/2,n-3} \leq \frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)} \leq t_{\alpha/2,n-3}\right] = 1-\alpha$$

$$\text{or, } P\left[\hat{\beta}_0 - t_{\alpha/2,n-3} \cdot SE(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2,n-3}\right] = 1-\alpha$$

when $\alpha = 0.05$, $P[\hat{\beta}_0 - t_{0.025,7} \times 23.570 \leq \beta_0 \leq \hat{\beta}_0 + t_{0.025,7} \times 23.570] = 1 - 0.05 = 0.95.$

or, $P[111.70 - 2.365 \times 27.570 \leq \beta_0 \leq 111.70 + 2.365 \times 23.570] = 0.95$

or, $P[55.957 \leq \beta_0 \leq 166.743] = 0.95$

$\therefore$ 95% confidence intervals of β_0 are 55.957 and 166.743

Similarly, when $\alpha = 0.01$, then 100 (1 – α)% = 99%.

So, 99% confidence intervals of β_0 would be

$$\hat{\beta}_0 \pm t_{\alpha/2, n-3} SE(\hat{\beta}_0)$$

or, $\hat{\beta}_0 \pm t_{0.005, 7} SE(\hat{\beta}_0)$

or, $111.70 \pm 3.499 \times 23.570$

or, 111.70 ± 82.4714

or, 29.2286 and 194.1714

i.e., 29.2286 and 194.1714

So, 99% confidence intervals of β_0 are 29.2286 and 194.1714

(vii) To test the null hypothesis $H_0 : \beta_1 = 0$ against the alternative $H_0 : \beta_1 \neq 0$ the test statistic under $H_0 : \beta_1 = 0$ would be t (observed) $= \dfrac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \sim t_{n-3}$

$$\text{Here } t \text{ (observed)} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{-7.19}{2.5592} = -2.8094$$

Now, $H_0 : \beta_1 = 0$ will be accepted if for the given sample $-t_{\alpha/2, n-3} \leq t$ (observed) $\leq t_{\alpha/2, n-3}$ and will be rejected otherwise.

When $\alpha = 0.05$, $t_{\alpha/2, n-3} = t_{0.025, (10-3)} = t_{0.025, 7} = 2.365$

and when $\alpha = 0.01$, $t_{\alpha/2, n-3} = t_{0.025, 7} = 3.499$.

Here we see that t (observed) $= -2.8094$ does not lie in the interval -2.365 and 2.365 and hence $H_0 : \beta_1 = 0$ is rejected at 5% level of significance. But t (observed) $= -2.8094$ lies in the interval -3.499 and 3.499 and hence $H_0 : \beta_1 = 0$ is accepted at 1% level of significance.

100 (1 − α)% confidence limits to β_1 would be :

$$\hat{\beta}_1 \pm t_{\alpha/2, n-3} \cdot SE(\hat{\beta}_1)$$

when $\alpha = 0.05$, then 95% confidence limits to β_1 would be :

$\hat{\beta}_1 \pm t_{0.025, 7} \cdot SE(\hat{\beta}_1)$

or, $-7.19 \pm 2.365 \times 2.5592$

or, -7.19 ± 6.0525 or, -13.245 and -1.135.

So, 95% confidence limits to β_1 are -13.245 and -1.135.

Similarly, when $\alpha = 0.01$, then 99% confidence limits to β_1 would be :

$\hat{\beta}_1 \pm t_{0.025, 7} \cdot SE(\hat{\beta}_1)$

or, $-7.19 \pm 3.499 \times 2.5592$

or, -7.19 ± 8.9546 or -16.1446 and 1.7646

So, 99% confidence limits to β_1 are -16.1446 and 1.7646

(viii) To test the null hypothesis $H_0 : \beta_2 = 0$ against the alternative $H_1 : \beta_2 \neq 0$ the appropriate test statistic under $H_0 : \beta_2 = 0$ would be

$$t = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} \sim t_{n-3}.$$

$$\text{Now } t \text{ (observed)} = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} = \frac{0.0143}{0.0111} = 1.2882.$$

$H_0 : \beta_2 = 0$ will be accepted if for the given sample $-t_{\alpha/2,n-3} \le t$ (observed) $\le t_{\alpha/2,n-3}$ and will be rejected otherwise.

When $\alpha = 0.05$, $t_{\alpha/2,n-3} = t_{0.025,7} = 2.365$

and when $\alpha = 0.01$, $t_{\alpha/2,n-3} = t_{0.005,7} = 3.499$

$$\text{Here we see that } t \text{ (observed)} = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} = 1.2882$$

which lies within both the intervals –2.365 to 2.365 and –3.499 to 3.449. So, $H_0 : \beta_2 = 0$ is accepted both at 5% and 1% levels of significance and hence insignificant.

Now 100 (1 – α)% confidence intervals of β_2 would be

$$\hat{\beta}_2 \pm t_{\alpha/2,n-3} \cdot SE(\hat{\beta}_2)$$

when α = 0.05, then 95% confidence intervals of β_2 would be :

$$\hat{\beta}_2 \pm t_{0.025,7} \cdot SE(\hat{\beta}_2)$$

or, $0.0143 \pm 2.365 \times 0.011$

or, 0.0143 ± 0.0262 or, -0.0077 and 0.0405.

So, 95% confidence intervals of β_2 are –0.0077 and 0.0405.

When α = 0.01, then 100 (1 – α)% = 99% confidence intervals of β_2 would be :

$$\hat{\beta}_2 \pm t_{0.005,7} \cdot SE(\hat{\beta}_2)$$

or, $0.0143 \pm 3.499 \times 0.0111$

or, 0.0143 ± 0.0388 or, -0.0245 and 0.0531

So, 99% confidence intervals of β_2 are –0.0245 and 0.0531

(ix) To test the null hypothesis $H_0 : \beta_1 = \beta_2$ against the alternative $H_1 : \beta_1 \neq \beta_2$ the appropriate test statistic under $H_0 : \beta_1 = \beta_2$ would be

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{SE(\hat{\beta}_1 - \hat{\beta}_2)} \sim t_{n-3}$$

$$\text{Now } t \text{ (observed)} = \frac{\hat{\beta}_1 - \hat{\beta}_2}{SE(\hat{\beta}_1 - \hat{\beta}_2)}$$

Since $\hat{\beta}_1 = -7.19$, $\hat{\beta}_2 = 0.0143$, $\text{var}(\hat{\beta}_1) = 6.55$, $\text{var}(\hat{\beta}_2) = 0.000124$ and

$$SE(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{\text{var}(\hat{\beta}_1) + \text{var}(\hat{\beta}_2) - 2\,\text{cov}(\hat{\beta}_1, \hat{\beta}_2)}$$

$$= \sqrt{6.55 + 0.000124 - 2 \times 0.0245}$$

$$= \sqrt{6.55 + 0.000124 - 0.049} = \sqrt{6.5011} = 2.549$$

$\therefore t$ (observed) $= \dfrac{\hat{\beta}_1-\hat{\beta}_2}{SE(\hat{\beta}_1-\hat{\beta}_2)} = \dfrac{-7.19-0.0143}{2.549} = \dfrac{-7.2043}{2.549} = -2.8263$

$\therefore t$ (observed) $= -2.8263$

Now, $H_0 : \beta_1 = \beta_2$ will be accepted if for the given sample $-t_{\alpha/2,n-3} \le t \le t_{\alpha/2,n-3}$ and will be rejected otherwise.

When $\alpha = 0.05$, $t_{\alpha/2,n-3} = t_{0.025,7} = 2.365$

and when $\alpha = 0.01$, $t_{\alpha/2,n-3} = t_{0.005,7} = 3.499$

Here we see that t (observed) $= -2.8263$ does not lie in the interval -2.365 to 2.365 and hence $H_0 : \beta_1 = \beta_2$ is rejected and $H_1 : \beta_1 \ne \beta_2$ is accepted at 5% level of significance.

But we see that t (observed) $= -2.8263$ lies in the interval -3.499 to 3.499 and hence the null hypothesis $H_0 : \beta_1 = \beta_2$ is accepted at 1% level of significance.

Now 100 (1 – α)% confidence intervals of $(\beta_1 - \beta_2)$ would be :

$$(\hat{\beta}_1-\hat{\beta}_2) \pm t_{\alpha/2,n-3} \cdot SE(\hat{\beta}_1-\hat{\beta}_2)$$

when α = 0.05, 100 (1 – α)% = 95% confidence intervals of $(\beta_1 - \beta_2)$ would be :

$$(\hat{\beta}_1-\hat{\beta}_2) \pm t_{0.025,7} \cdot SE(\hat{\beta}_1-\hat{\beta}_2)$$

or, $(-7.19 - 0.0143) \pm 2.365 \times 2.549$

or, -7.2043 ± 6.0283 or, -13.2326 and -1.176.

So, 95% confidence intervals of $(\beta_1 - \beta_2)$ are $-$ 13.2326 and -1.176.

When α = 0.01, then 100 (1 – α)% = 99% confidence intervals of $(\beta_1 - \beta_2)$ would be :

$$(\hat{\beta}_1-\hat{\beta}_2) \pm t_{\alpha/2,n-3} \cdot SE(\hat{\beta}_1-\hat{\beta}_2)$$

or, $(-7.19 - 0.0173) \pm t_{0.005,7} \times 2.549$

or, $-7.2043 \pm 3.499 \times 2.549$

or, -7.2043 ± 8.9189 or, -16.1232 and 1.7146

So, 99% confidence intervals of $(\beta_1 - \beta_2)$ are -16.1232 and 1.7146

(x) We have to construct 95% and 99% confidence intervals of σ_u^2.

We know that 100(1 – α)% confidence intervals of σ_u^2 would be given by

$$(n-3)\frac{\hat{\sigma}_u^2}{\chi^2_{\alpha/2,n-3}} \text{ and } (n-3)\frac{\hat{\sigma}_u^2}{\chi^2_{1-\alpha/2,n-3}}$$ where χ^2 values are taken from the table with d.f = $(n - 3)$

When $\alpha = 0.05$, $\chi^2_{\alpha/2,n-3} = \chi^2_{0.025,7} = 16.013$

and $\chi^2_{1-\alpha/2,n-3} = \chi^2_{0.0975,7} = 1.690$

So, 95% confidence intervals of σ_u^2 would be

$$(n-3)\frac{\hat{\sigma}_u^2}{\chi^2_{\alpha/2,n-3}} = (10-3)\times\frac{52.24}{16.013} = \frac{7\times 52.24}{16.013} = 22.836$$

and $$(n-3)\frac{\hat{\sigma}_u^2}{\chi^2_{1-\alpha/2,n-3}} = (10-3)\times\frac{52.24}{1.690} = \frac{7\times 52.24}{1.690} = 216.378$$

Again when $\alpha = 0.01$, $\chi^2_{\alpha/2,n-3} = \chi^2_{0.005,7} = 20.278$

and $\chi^2_{1-\alpha/2,n-3} = \chi^2_{0.995,7} = 0.989$

So, 99% confidence intervals of σ_u^2 would be :

$$(n-3)\frac{\hat{\sigma}_u^2}{\chi^2_{\alpha/2,n-3}} = (10-3)\times\frac{52.24}{20.278} = \frac{7\times 52.24}{20.278} = 18.033$$

and $$(n-3)\frac{\hat{\sigma}_u^2}{\chi^2_{1-\alpha/2,n-3}} = (10-3)\times\frac{52.24}{\chi^2_{0.995,7}} = \frac{7\times 52.24}{0.989} = 369.747$$

3.10. Analysis of Variance (ANOVA) in a Multiple Linear (Three-Variable) Regression Model

Yet another item that is often presented in connection with the three variable linear regression model is the **analysis of variance.** This is the break down of the total sum of squares (TSS) into explained sum of squares (ESS) and the residual sum of squares (RSS).

The estimated three variable linear regression line (where the regression equation is $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, i = 1, 2, ..., n$) is given by $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$ and $\bar{Y} = \hat{\beta}_0 + \beta_1 \bar{X}_1 + \beta_2 \bar{X}_2$

Taking deviations from mean we have

$$\hat{Y}_i - \bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} - \hat{\beta}_0 - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$$

or, $\hat{y}_i = \hat{\beta}_1 (X_{1i} - \bar{X}_1) + \hat{\beta}_2 (X_{2i} - \bar{X}_2)$

or, $\hat{y}_i = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$, where $x_{1i} = X_{1i} - \bar{X}_1$ and $x_{2i} = X_{2i} - \bar{X}_2$.

Now, error of estimate, $e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}$ $\therefore y_i = \hat{y}_i + e_i$

Now, $\Sigma y_i^2 = \Sigma(\hat{y}_i + e_i)^2 = \Sigma\hat{y}_i^2 + 2\Sigma\hat{y}_i \cdot e_i + \Sigma e_i^2$

$\therefore \Sigma y_i^2 = \Sigma\hat{y}_i^2 + \Sigma e_i^2$ as $\Sigma\hat{y}_i e_i = 0$ by assumption.

i.e., TSS = ESS + RSS

with d.f : $(n-1) = (K-1) + (n-K)$ where K = number of parameters. Here $K = 3$ as there are three parameters β_0, β_1 and β_2 in a three variable linear regression model.

Here, $\text{ESS} = \Sigma\hat{y}_i^2 = \hat{\beta}_1\Sigma x_{1i}y_i + \hat{\beta}_2\Sigma x_{2i}y_i$

Thus we see that total variations are split into explained (by explanatory variables) and unexplained (error terms) variations against between and within variations in case of analysis of variance procedure. This suggests that we can compile an analysis of variance type of table for the regression analysis also in order to judge the overall significance of the regression results.

ANOVA TABLE

Table 3.4

Source of variation	Sum of squares	Degress of freedom	Mean sum of squares	F	
				Observed	Tabulated
Explained (between)	$ESS = \Sigma\hat{y}_i^2$ $= \hat{\beta}_1\Sigma x_{1i}y_i$ $+ \hat{\beta}_2\Sigma x_{2i}y_i$	$K-1$	$\frac{ESS}{K-1}$ $= MSE$	$F = \frac{MSE}{MSR}$ with d.f. $= (K-1), (n-K)$	
Residual (within)	$RSS = \Sigma e_i^2$	$n-K$	$\frac{RSS}{n-K}$ $= MSR$		
Total	$TSS = \Sigma y_i^2$	$n-1$			

In case of a three variable linear regression model there are three parameters and hence $K = 3$.

The test aims at finding out whether the explanatory variables (X_1 and X_2) do actually have any significant influence on the dependent variable Y. Formally the test of the overall significance of the regression implies testing the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ against the alternative hypothesis H_1 : not all $\beta_i{}^s$ are zero. We may use the test statistic

$$F^* = \frac{MSE}{MSR} = \frac{\Sigma\hat{y}_i^2/(K-1)}{\Sigma e_i^2/(n-K)}$$

$$= \frac{\Sigma\hat{y}_i^2/(3-1)}{\Sigma e_i^2/(n-3)} = \frac{\Sigma\hat{y}_i^2/2}{\Sigma e_i^2/(n-3)} \text{ with d.f.} = 2,\ n-3 \text{ (Here } K = 3)$$

Now we have to compare $F^*_{2,n-3}$ with the table value of F with d.f. = 2, $n-3$. If it is found that $F^* > F_{\alpha;2,n-3}$ (Table value) we reject the null hypothesis at 100 α% level of significance (α = 0.01 or 0.05 usually), i.e. we accept that the regression is significant and not all β_i's are zero.

If $F^* < F_{\alpha;2,n-3}$ we accept the null hypothesis that is we accept that the overall regression is not significant.

Note : Relation between R^2 and F

There is an intimate relationship between the coefficient of multiple determination R^2 and the F test used in the analysis of variance. Assuming the normal distribution for the disturbances u_i and the null hypothesis that $H_0 : \beta_1 = \beta_2 = 0$, we see that

$$F^* = \frac{ESS/(K-1)}{RSS/(n-K)} = \frac{ESS/2}{RSS/(n-3)}$$

is distributed as the F distribution with d.f = 2 and $(n-3)$ [Here $K = 3$ as there are three paramaters including the constant intercept term β_0] in a three variable linear regression model.

$$\text{Now we can write } F = F^* = \frac{ESS/(K-1)}{RSS/(n-K)}$$

$$= \frac{n-K}{K-1} \cdot \frac{ESS}{RSS} = \frac{n-K}{K-1} \cdot \frac{ESS}{TSS-ESS}$$

$$= \frac{n-K}{K-1} \cdot \frac{ESS/TSS}{1-\frac{ESS}{TSS}} = \frac{n-K}{K-1} \times \frac{R^2}{1-R^2} = \frac{R^2/(K-1)}{(1-R^2)/(n-K)}$$

$$\therefore F^* = \frac{R^2}{(K-1)} \Big/ \frac{(1-R^2)}{(n-K)}.$$

It should be noted that here K = number of parameters in the linear regression model. Here $K = 3$ when there are two explanatory variables. When $R^2 = 0$, $F^* = 0$. The larger the R^2, the greater the F^* value.

In the limit, when $R^2 = 1$, F is infinite. Thus, the F test, which is a measure of the over all significance of the estimated regression, is also a test of significance of R^2. In other words, testing the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ is equivalent to test the null hypothesis that (population) R^2 is zero. The ANOVA table can also be written/ expressed in terms of R^2 as shown below :

ANOVA TABLE in terms of R²

Table 3.5

Source of variation	Sum of squares (SS)	Degress of freedom (d.f)	Mean sum of squares (MS)	F	
				Observed	Tabulated
Explained (between)	$ESS = \Sigma\hat{y}_i^2$ $= R^2\Sigma y_i^2$	$K - 1$	$ESS/(K-1)$ $= MSE$	$F = \frac{MSE}{MSR}$ with d.f. $= (K-1), (n-K)$	
Residual (within)	$RSS = \Sigma e_i^2 =$ $= (1-R^2)\Sigma y_i^2$	$n - K$	$RSS/(n-K)$ $= MSR$		
Total	$TSS = \Sigma y_i^2$	$n - 1$	—	—	—

In case of three variable linear regression model there are three paramaters and hence we put $K = 3$.

Example 3.9. Following example 3.8. test the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ against the alternative $H_1 : \beta_1$ and β_2 are not zero.

Solution : In testing the null hypothesis $H_0 : \beta_1 = 0$ against $H_0 : \beta_1 \neq 0$ and $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$ we used t statistic $\left(t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \text{ and } t = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)}\right)$ but here we have to test $H_0 : \beta_1 = \beta_2 = 0$ (jointly) against $H_0 : \beta_1$ and β_2 are not zero.

In this case we have to use F statistic and ANOVA. The calculations are shown in the following ANOVA Table.

ANOVA TABLE relating to demand for a commodity

Table 3.6

Source of variation	Sum of squares (SS)	Degress of freedom (d.f)	Mean sum of squares (MS)	F	
				Observed	Tabulated
Explained (between)	$ESS = \Sigma\hat{y}_i^2$ $=3086.5$	$K - 1 = 2$	$MSE = \frac{ESS}{K-1}$ $=1543.25$	$F = F^*$ $= \frac{MSE}{MSR}$	$F_{0.05;2,7}$ $= 7.74$
Residual (within)	$RSS = \Sigma e_i^2 =$ $=363.5$	$n - K = 7$	$MSR = \frac{RSS}{n-K}$ $=51.92$	with d.f. $= (K-1), (n-K)$ $= 29.72$	$F_{0.01;2,7}$ $= 9.55$
Total	$\Sigma y_i^2 = 3450$	$n - 1 = 9$	—	—	—

Here the sample size, $n = 10$, number of paramaters, $K = 3$.

$\therefore K - 1 = 2$ and $n - K = 10 - 3 = 7$.

From Example 3.8. we have obtained the results

$$ESS = \Sigma\hat{y}_i^2 = \hat{\beta}_1 \Sigma x_{1i} y_i + \hat{\beta}_2 \Sigma x_{2i} y_i = 3086.5$$

$$RSS = \Sigma e_i^2 = 363.5 \text{ and } TSS = \Sigma y_i^2 = 3450$$

Now the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ will be rejected if for the given sample

$$F = F^* \text{ (observed)} = \frac{MSE}{MSR} \text{ [with d.f } (K-1) = 2 \text{ and } (n-K) = 7]$$

is greater than the table value of F with d.f : $(K - 1) = 2$ and $(n - K) = 7$. From the table value we see that $F_{0.05;2,7} = 7.74$ and $F_{0.01;\,2,7} = 9.55$

Here we see that F (observed) $= F^* = 29.72$ and $F_{0.05\,;2,7} = 7.74$

$\therefore F^* > F_{0.5;2.7}$.

So, at 5% level of significance the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ will be rejected for the given sample.

We also see that $F^* = 29.72 > F_{0.01;2,7} = 9.55$. This means that at 1% level of significance the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ will be rejected for the given sample.

Thus both at 1% and 5% levels of significance we may claim that the slope coefficients of the regression equation are not zero.

It should be noted that we can also construct the ANOVA table in terms of R^2. Following Example 3.8 we are showing the ANOVA table below in tersm of R^2.

ANOVA TABLE in terms of R^2

Table 3.7

Sourceof variation	Sum of squares (SS)	Degress of freedom (d.f)	Mean sum of squares (MS)	F Observed	F Tabulated
Explained (between)	$ESS = R^2 \cdot \Sigma y_i^2$ $= 0.894 \times 3450$ $= 3084.30$	$K - 1 = 2$	$ESS/(K-1)$ $= MSE$ $= \dfrac{3084.30}{2}$ $= 1542.15$	$F = F^*$ $= \dfrac{MSE}{MSR}$ $= \dfrac{1542.15}{52.242}$ $= 29.52$	$F_{0.05;2,7}$ $= 7.74$
Residual (within)	RSS $= (1-R^2)\,\Sigma y_i^2$ $= 0.106 \times 3450$ $= 365.70$	$n - K = 7$	$RSS/(n-K)$ $= MSR$ $= \dfrac{365.70}{7}$ $= 52.242$		$F_{0.01;2,7}$ $= 9.55$
Total	TSS $\Sigma y_i^2 = 3450$	$n - 1 = 9$	—	—	—

From Example 3.8, we have seen that $n = 10$, $K = 3$, $\Sigma y_i^2 = 3450$ and $R^2 = 0.894$. Here also we see that $F = F^* = 29.52 > F_{0.05;2,7} = 7.74$ and $F^* = 29.52 > F_{0.01,2,7} = 9.55$.

Thus the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ is rejected both at 1% and 5% levels of significance.

3.11. The Cobb-Douglas Production Function : More on Functional Form

In Section 2.18 we showed how with appropriate transformations we can convert non-linear relationships into linear ones so that we can work within the framework of classical linear regression model. We consider the Cobb-Douglas Production function which shows a three variable non-linear relation. The Cobb-Douglas Production function, in its stochastic form, may be expressed as

$$Y_i = \beta_0 X_{1i}^{\beta_1} X_{2i}^{\beta_2} U_i$$

where Y = output, X_1 = labour input, X_2 = capital input, U = Stochastic disturbance term, β_0 = constant technological parameter.

Taking Log on both sides of the above non-linear equation we obtain :

$$\text{Log } Y_i = \log \beta_0 + \beta_1 \log X_{1i} + \beta_2 \log X_{2i} + \log U_i$$

or, $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i,\ u_i \sim N(0, \sigma_u^2)$

This is a linear equation in terms of logarithms, where $y_i = \log Y_i$, $x_{1i} = \log X_{1i}$, $x_{2i} = \log X_{2i}$, $u_i = \log U_i$ and $\alpha = \log\beta_0$ and $\beta_0 =$ Anti-log of $\log \beta_0$.

We can now apply the usual rule for determining the OLS estimators of the regression parameters. The usual properties of the Cobb-Douglas production function are :

(i) β_1 is the (partial) elasticity of output with respect of labour input i.e. $\dfrac{\partial \log Y_i}{\partial \log X_{1i}} = \beta_1$.

(ii) Likewise, β_2 is the (partial) elasticity of output with respect to capital input, i.e. $\dfrac{\partial \log Y_i}{\partial \log X_{2i}} = \beta_2$.

(iii) The sum $(\beta_1 + \beta_2)$ gives information about the returns to scale. The function displays IRS, CRS and DRS according as $\beta_1 + \beta_2) \gtreqless 1$.

Example 3.10. A production function is specified as $Y_i = \beta_0 X_{1i}^{\beta_1} \cdot X_{2i}^{\beta_2} U_i$ where $i = 1, 2, \ldots, n$

$Y =$ output, $X_1 =$ labour input, $X_2 =$ capital input, $U =$ Stochastic disturbance term, $n =$ sample size. The corresponding Log-Linear form of the production function is given as

$$\text{Log } Y_i = \log \beta_0 + \beta_1 \log X_{1i} + \beta_2 \log X_{2i} + \log U_i$$

$$\text{or, } y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i,\ u_i \sim N(0, \sigma_u^2)$$

On the basis of a sample size of 23 the following results are given : $\hat{\alpha} = 4.0$, $\hat{\beta}_1 = 0.7$, $\hat{\beta}_2 = 0.2$, $RSS = 1.4$, $TSS = 10$, $\text{var}(\hat{\alpha}) = 0.6084$, $\text{var}(\hat{\beta}_1) = \text{var}(\hat{\beta}_2) = 0.0105$.

(i) Write the estimated regression equation.

(ii) Find the value of R^2.

(iii) Find $SE(\hat{\alpha})$, $SE(\hat{\beta}_1)$, and $SE(\hat{\beta}_2)$

(iv) Find $\hat{\sigma}_u^2$.

(v) Find the 95% confidence intervals for α, β_1, β_2 and σ_u^2.

(vi) Test the hypothesis $\beta_1 = 1.0$ and $\beta_2 = 0$ separately at the 5% significance level.

Solution : (i) The estimated regression equation can be written as

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$$

$$\text{or, } \quad \hat{y}_i = 4.0 + 0.7 x_{1i} + 0.2 x_{2i}$$

(ii) The value of multiple coefficient of determination is given by $R^2 = \frac{ESS}{TSS} = \frac{8.6}{10} = 0.86$

where $ESS = TSS - RSS = 10 - 1.4 = 8.6$ $\therefore R^2 = 0.86$.

(iii) We know that $SE(\hat{\alpha}) = \sqrt{\text{var}(\hat{\alpha})} = \sqrt{0.6084} = 0.78$

similarly, $SE(\hat{\beta}_1) = \sqrt{\text{var}(\hat{\beta}_1)} = \sqrt{0.0105} = 0.102$

and $SE(\hat{\beta}_2) = \sqrt{\text{var}(\hat{\beta}_2)} = \sqrt{0.0105} = 0.102$.

(iv) We have to find out the value of OLS estimator of σ_u^2 i.e., $\hat{\sigma}_u^2$.

We know that $\hat{\sigma}_u^2 = \frac{\Sigma e_i^2}{n-3} = \frac{RSS}{n-3}$

Here $RSS = 1.4$ and $n = 23$

$\therefore \hat{\sigma}_u^2 = \frac{RSS}{n-3} = \frac{1.4}{23-3} = \frac{1.4}{20} = 0.07.$

(v) We have to find out the 95% confidence intervals for α, β_1, β_2 and σ_u^2. Using the 't' distribution with d.f = $(n - 3) = (23 - 3) = 20$, we can get the 95% confidence intervals for α, β_1, and $\hat{\beta}_2$ as :

For α : $\hat{\alpha} \pm t_{0.025,20} SE(\hat{\alpha}) = 4.0 \pm 2.086 \times 0.78 = 4.0 \pm 1.63$

$= (2.37, 5.63)$ [$\because t_{\alpha/2, n-3} = t_{0.025,20} = 2.086$, as $n = 23$, $\alpha = 0.05$]

For β₁ : $\hat{\beta}_1 \pm t_{0.025,20} \, SE(\hat{\beta}_1) = 0.7 \pm 2.086 \times 0.102$

$= 0.7 \pm 0.21 = (0.49, 0.91)$

and **for β₂ :** $\hat{\beta}_2 \pm t_{0.025,20} \, SE(\hat{\beta}_2) = 0.2 \pm 2.086 \times 0.102$

$= 0.2 \pm 0.21 = (-0.01, 0.41)$

Again, 95% confidence intervals for σ_u^2 would be

$$P\left[(n-3)\frac{\hat{\sigma}_u^2}{\chi^2_{\alpha/2, n-3}} \le \sigma_u^2 \le (n-3)\frac{\hat{\sigma}_u^2}{\chi^2_{1-\alpha/2, n-3}}\right] = 1 - \alpha$$

when $\alpha = 0.05$, $n = 23$, $\hat{\sigma}_u^2 = 0.07$

$$P\left[20 \times \frac{0.07}{\chi^2_{0.025;20}} \le \sigma_u^2 \le 20 \times \frac{0.07}{\chi^2_{0.975;20}}\right] = 1 - 0.05 = 0.95$$

or, $P\left[\frac{1.4}{34.170} \le \sigma_u^2 \le \frac{1.4}{9.591}\right] = 0.95$

or, $P\left[0.041 \le \sigma_u^2 \le 0.146\right] = 0.95$

$\therefore$ 95% confidence intervals for σ_u^2 are 0.041 and 0.146.

(vi) To test the null hypothesis $H_0 : \beta_1 = 1.0$ against the alternative $H_0 : \beta_1 \neq 1$, the appropriate test statistic under the $H_0 : \beta_1 = 1.0$ would be

$$t \text{ (observed)} = \frac{\hat{\beta}_1 - 1}{SE(\hat{\beta}_1)} \sim t_{n-3}$$

$$\therefore t \text{ (observed)} = \frac{\hat{\beta}_1 - 1}{SE(\hat{\beta}_1)} = \frac{0.7-1}{0.102} = \frac{-0.3}{0.102} = -2.941$$

Now $H_0 : \beta_1 = 1.0$ will be accepted if for the given sample $-t_{\alpha/2, n-3} \leq t$ (observed) $\leq t_{\alpha/2, n-3}$ and will be rejected otherwise.

When $\alpha = 0.05$, $t_{\alpha/2, n-3} = t_{0.025, 20} = 2.086$

Here we see that t (observed) $= -2.941$

does not lie in the interval $-t_{0.025,20} = -2.086$ and $t_{0.025,20} = 2.086$ and hence the null hypothesis $H_0 : \beta_1 = 1.0$ is rejected at 5% level of significance.

Again to test the null hypothesis $H_0 : \beta_2 = 0$ against the alternative $H_1 : \beta_2 \neq 0$, the appropriate test statistic under $H_0 : \beta_2 = 0$ would be

$$t \text{ (observed)} = \frac{\hat{\beta}_2 - 0}{SE(\hat{\beta}_2)} \sim t_{n-3}$$

$$\text{Here } t \text{ (observed)} = \frac{0.2}{0.102} = 1.960$$

Now, $H_0 : \beta_2 = 0$ will be accepted if for the given sample $-t_{\alpha/2, n-3} \leq t$ (observed) $\leq t_{\alpha/2, n-3}$ and will be rejected otherwise.

When $\alpha = 0.05$, $t_{\alpha/2, n-3} = t_{0.025, 20} = 2.086$

Here we see that t (observed) $= 1.960$ lies in the interval -2.086 and 2.086 and hence $H_0 : \beta_2 = 0$ is accepted at 5% level of significance.

3.12. Prediction / Forecasting in the Multiple (Three-Variable) Regression Model

The formulas for prediction in multiple regression are similar to those in the case of simple (two variable linear regression) regression, except that to compute the standard error of the predicted value we need the variances and covariances of all regression parameters. Here we will present the expression for the standard error in the case of two explanatory variables.

Let the estimated regression equation be,

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 .$$

Now consider the prediction of the value Y_0 of Y, given values X_{10} of X_1 and X_{20} of X_2, respectively. These could be values at some future date.

Then we have $Y_0 = \beta_0 + \beta_1 X_{10} + \beta_2 X_{20} + u_0$

and $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_{10} + \hat{\beta}_2 X_{20}$.

The prediction error is $e_0 = \hat{Y}_0 - Y_0$

$$= (\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)X_{10} + (\hat{\beta}_2 - \beta_2)X_{20} - u_0$$

Since $E(\hat{\beta}_0 - \beta_0)$, $E(\hat{\beta}_1 - \beta_1)$, $E(\hat{\beta}_2 - \beta_2)$ and $E\,(u_0)$ are all equal to zero, we have

$E(\hat{Y}_0 - Y_0) = 0$ or, $E(\hat{Y}_0) = E(Y_0) = Y_0$ $\therefore E(\hat{Y}_0) = Y_0$

Thus, the predictor $\hat{Y}_0$ is unbiased.

The variance of the predictor / prediction error is given by

[where $\text{var}(\hat{Y}_0) = E[\hat{Y}_0 - E(\hat{Y}_0)]^2 = E[\hat{Y}_0 - Y_0]^2 = E[e_0]^2 = \text{var}\,(e_0)$]

$$\text{var}(\hat{Y}_0) = \text{var}(e_0) = \sigma_u^2\left(1 + \frac{1}{n}\right) + \left(X_{10} - \bar{X}_1\right)^2 \cdot \text{var}(\hat{\beta}_1)$$

$$+ 2(X_{10} - \bar{X}_1)(X_{20} - \bar{X}_2)\,\text{cov}(\hat{\beta}_1, \hat{\beta}_2) + (X_{20} - \bar{X}_2)^2 \cdot \text{var}(\hat{\beta}_2)$$

where $\text{var}(\hat{\beta}_1) = \dfrac{\sigma_u^2 \Sigma x_{2i}^2}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2}$, $\text{var}(\hat{\beta}_2) = \dfrac{\sigma_u^2 \Sigma x_{1i}^2}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2}$

and $\text{cov}(\hat{\beta}_1, \hat{\beta}_2) = \dfrac{-\sigma_u^2 \Sigma x_{1i} x_{2i}}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2}$

when $x_{1i} = X_{1i} - \bar{X}_1$, $x_{2i} = X_{2i} - \bar{X}_2$.

When σ_u^2 is not known, it is replaced by its unbiased estimator $\hat{\sigma}_u^2 = \Sigma e_i^2 / (n-3)$, n = sample size. Thus 100(1 – α) % confidence interval for the prediction would be $\hat{Y}_0 \pm t_{\alpha/2, n-3} \cdot SE(\hat{Y}_0)$ where $SE(\hat{Y}_0) = +\sqrt{\text{var}(\hat{Y}_0)}$.

Example 3.11. Following Example 3.8,

(i) using the estimated regression, $\hat{Y}_i = 111.70 - 7.19X_{1i} + 0.0143X_{2i}$, find the prediction of Y for $X_{10} = 10$ and $X_{20} = 1400$.

(ii) Using the results of $\text{var}(\hat{\beta}_1)$, $\text{var}(\hat{\beta}_2)$, $\text{cov}(\hat{\beta}_1, \hat{\beta}_2)$ and $\hat{\sigma}_u^2$ derived in Example 3.8, estimate the variance of the prediction error and the standard error of the prediction error.

(iii) Find 95% confidence interval for the prediction.

Solution :

(i) The prediction of Y can be obtained from the estimated regression equation,

$\hat{Y}_i = 111.70 - 7.19X_{1i} + 0.0134X_{2i}$.

Here we put $X_{10} = 10$ and $X_{20} = 1400$ and get

$\hat{Y}_0 = 111.70 - 7.19 \times 10 + 0.0134 \times 1400 = 58.56$

∴ The predicted value of Y is $\hat{Y}_0 = 58.56$ when $X_{10} = 10$ and $X_{20} = 1400$.

(ii) From example 3.8 we obtain, $\text{var}\,(\hat{\beta}_1) = 6.35$, $\text{var}\,(\hat{\beta}_2) = 0.000124$, $\text{cov}\,(\hat{\beta}_1, \hat{\beta}_2) =$

0.0245, $\hat{\sigma}_u^2 = 52.24$, $\bar{X}_1 = 6$, $\bar{X}_2 = 800$ and $n = 10$. We now put these values in the expression of the variance of the prediction,

$$\text{var}(\hat{Y}_0) = \text{var}(e_0) = \hat{\sigma}_u^2\left(1+\frac{1}{n}\right)+(X_{10}-\bar{X}_1)^2\,\text{var}(\hat{\beta}_1)$$

$$+\ 2(X_{10}-\bar{X}_1)(X_{20}-\bar{X}_2)\,\text{cov}(\hat{\beta}_1,\hat{\beta}_2)+(X_{20}-\bar{X}_2)^2\,\text{var}(\hat{\beta}_2) \text{ and get}$$

$$\text{var}(\hat{Y}_0) = 52.24\left(1+\frac{1}{10}\right)+(10-6)^2\times 6.55$$

$$+\,2(10-6)(1400-800)\times 0.0245+(1400-800)^2\times 0.000124$$

[Here $X_{10} = 10$ and $X_{20} = 1400$, which are given]

$$= 52.24\times 1.1+(4)^2\times 6.55+2\times 4\times 600\times 0.0245+(600)^2\times 0.000124$$

$$= 57.464+104.8+117.6+44.64$$

$$= 324.504.$$

$$\therefore \text{var}(\hat{Y}_0) = 324.504$$

and $SE(\hat{Y}_0) = \sqrt{\text{var}(\hat{Y}_0)} = \sqrt{324.504} = 18.014$

∴ Variance of the prediction is 324.504 and standard error of prediction is 18.014.

(iii) We know that $(1-\alpha)$ 100% confidence interval for prediction would be, $\hat{Y}_0 \pm t_{\alpha/2,n-3}\cdot SE(\hat{Y}_0)$ when $\alpha = 0.05$, $t_{\alpha/2,n-3} = t_{0.025\,;\,(10-3)} = t_{0.025,7} = 2.365$

∴ 95% confidence interval for prediction would be

$\hat{Y}_0 \pm t_{0.025,7}\cdot SE(\hat{Y}_0)$

or, $58.56 \pm 2.365\times 18.014$ or, 58.56 ± 42.60 or, (15.96, 101,16)

∴ 95% confidence interval for prediction would be 15.96 and 101.16.

3.13. Regression Analysis in Presence of Qualitative (Dummy) Variables

3.13.1. Meaning

In Section 1.10 we mentioned four types of variables that one generally encounters in empirical analysis. These are : ratio scale, interval scale, ordinal scale and nominal scale. The types of variables used in earlier sections were essentially ratio scale. In many cases we deal with models that may involve not only ratio scale variables but also nominal scale variables. Such variables are known as **indicator variables, categorical variables, qualitative variables**, or, **dummy variables (Binary variables).**

3.13.2. Nature of Dummy Variables

In regression analysis the dependent variable or regressand is frequently influenced not only by ratio scale variables (e.g., income, output, prices, costs, height, weight, temperature, etc.) but also by variables that are essentially qualitative or nominal scale, in nature, such as sex, race, colour, religion, nationality, geographical region, political upheavals and party affiliation. For example, holding all other factors constant, female workers are found to earn less than their male counterparts or non white workers are found to earn less than whites. This pattern may result from sex or racial discrimination, but whatever the reason, qualitative variables such as sex, race which seem to influence the regressand clearly should be included among the explanatory variables or the regressors.

Since such variables usually indicate the presence or absence of a 'quality' or an attribute, such as male or female, black or white etc. which are essentially nominal scale variables, one way we could 'quantify' such attributes is by constructing artificial variables that take on values of 1 or 0, i.e., 1 indicating the presence (or possession) of that attribute and 0 indicating the absence of that attribute. For example, 1 may indicate that a person is a male and 0 may designate a female ; or 1 may indicate that a person is literate, and 0 that the person is illiterate, and so on. Dummy variables can be incorporated in regression models just as easily as quantitative variables. As a matter of fact a regression model may contain regressors that are all exclusively dummy in character. Such models are called **Analysis of variance (ANOVA) models.**

3.13.3. Use of Dummy Variables

The dummy variables are used differently in the regression analysis. Dummy variables can be used for several purposes. We are explaining some of the uses of dummy variables in applied economic research.

(i) Dummy variables as proxies to Qualitative (Categorical) Factors

Dummy variables are commonly used as proxies for qualitative factors like profession, religion, sex, region etc. For example we consider a sample of family budgets from all regions of the country, rural and urban and we want to estimate the demand for tobacco manufactures as a function of income. It is known that town dwellers are heavier smokers than rural farmers. Thus, 'region' is an important explanatory factor in this case. We may represent this factor by a dummy variable to which we might arbitrarily assign the value 1 for a town dweller and 0 for a person living in a rural area. The demand function can be written as

$$D_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

where X_1 = income, and X_2 = dummy variable for region. We put $X_2 = 1$ for a town dweller ($\beta_2 > 0$) and $X_2 = 0$ for a person living in rural area.

(ii) Dummy variables as proxies of Numerical Factors

Dummy variables may be used as proxies for quantitiative factors, when no observations on these factors are available or when it is convenient to do so. For example, suppose we want to measure saving function $S = f(Y)$ from a cross section sample of consumers. Obviously, 'age' is an important explanatory factor of the consumption and savings patterns of a community, since people become more thrifty with age. Although 'age' is a quantitative factor, we may approximate it by a dummy variable. We may divide the selected consumers in two age groups :

Group 1 : People of 20–40 years of age

Group 2 : People of 40 years and over

On the assumption that people become more thirfty as they grow old, the dummy variable for 'age' may be assigned the value 0, if the person belongs to group 1 and the value 1 if the person belongs to group-2.

The saving function can be written in the form :

$$S_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

where S_i = saving, X_1 = income,

X_2=dummy variable for age $\begin{cases} =0 \text{ if the person belongs to group-} \\ =1 \text{ if the person belongs to group-} \end{cases}$

where $\beta_2 > 0$.

(iii) Dummy variables are used for measuring the shift of a function over time

A shift of a function implies that the constant intercept changes in different periods, while other coefficients remain constant. Such shifts may be taken into account by the introduction of a dummy variable in the function.

For example suppose that we have data on the consumption for an economy for the period 1900-1968. During this period the economy faced two World Wars (1914-1918 and 1940-1945) and a deep depression (1929-1933). The abnormal conditions prevailing in these years have caused a shift of the consumption function downwards, due to rationing, various controls and other factors. To capture this shift we may use a dummy variable say 'Z' which would assume the value 0 during the above 'abnormal' years and 1 in the other normal years. The consumption function takes the form :

$$C_t = \beta_0 + \beta_1 Y_t + \beta_2 Z_t + u_t, (\beta_2 > 0)$$

where C = consumption, Y = income,

Z = Dummy variable for the shift of the function.

For a normal year the estimated form of the consumption function would be :

$$\hat{C} = \hat{\beta}_0 + \hat{\beta}_1 Y + \hat{\beta}_2 = (\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 Y$$

and for an abnormal period it would be

$$\hat{C} = \hat{\beta}_0 + \hat{\beta}_1 Y$$

If we plot these two functions we can clearly see the shift of the consumption function during the abnormal (War and depression) years.

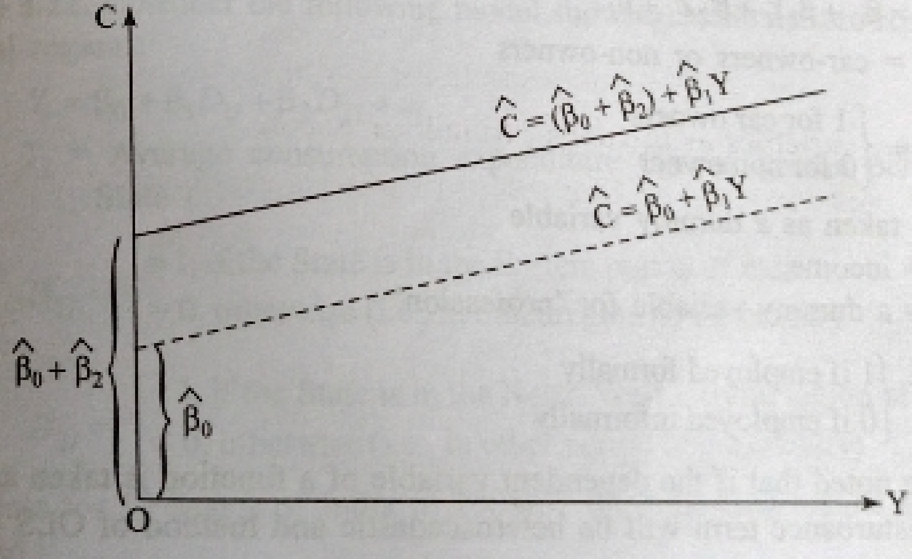

C
$\hat{C} = (\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 Y$
$\hat{C} = \hat{\beta}_0 + \hat{\beta}_1 Y$
$\hat{\beta}_0 + \hat{\beta}_2$
$\hat{\beta}_0$
O
Y

Fig. 3.1.

The slope of the consumption function (Fig. 3.1) i.e., *MPC* is assumed to be the same both in normal and abnormal periods and hence the two regression lines are parallel (only intercept changes, slope remaining the same).

(iv) Dummy variables are used for measuring the change of parameters (slopes) over time

It is usual that over long periods of time or in abnormal (War or depression) years not only do the functions shift (their constant intercept changes) but also their slope(s) may well be expected to change : elasticities and propensities change over time. The change in the parameter of a function may be captured by introducing appropriate dummy variables in the function.

We can write here the consumption function in the form :

$$C_t = \beta_0 + \beta_1 Y_t + \beta_2 Z_{1t} + \beta_3 Z_{2t} + u_t$$

where : C = consumption, Y = income

$$Z_1 = \text{Dummy variable} \begin{cases} = 0 \text{ for abnormal years} \\ = 1 \text{ for normal years} \end{cases}$$

$$Z_2 = YZ_1 = \text{Dummy variable} \begin{cases} 0 \text{ for abnormal years (when } Z_1 = 0) \\ Y \text{ for normal years(when } Z_1 = 1) \end{cases}$$

Consequently for a normal period the estimated consumption function would be given by

$\hat{C} = (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_3)Y$, while for an abnormal year the estimated function would be

$$\hat{C} = \hat{\beta}_1 Y$$

In this case both slope and intercept of the function will change.

(v) Dummy variables are used as proxies for the dependent variable

In some cases, the dependent variable of a function may be a dummy variable.

For example, suppose we want to measure the determinants of car-ownership from a cross-section sample. Some people will have cars while others will not. Suppose that the determinants of the ownership depend on income and profession.

The functional relation can be written in the form :

$$C_i = \beta_0 + \beta_1 Y_i + \beta_2 Z_i + u_i$$

where C = car-owners or non-owners

$$= \begin{cases} 1 \text{ for car owners} \\ 0 \text{ for non owner} \end{cases}$$

Here C is taken as a dummy variable

Y = income,

Z = a dummy variable for 'profession'

$$= \begin{cases} 1 \text{ if employed formally} \\ 0 \text{ if employed informally} \end{cases}$$

It should be noted that if the dependent variable of a function is taken as a dummy variable, the disturbance term will be heteroscedastic and method of OLS will not be appropriate there.

(vi) Dummy variables are used for seasonal adjustments of time series

One of the most common use of dummy variables is in removing seasonal variations in time series. For example, if we have quarterly data on retail sales, we should adjust for bulk purchases at Christmas and Easter before attempting to measure the influence of other factors on demand. The seasonal adjustment in this case can be estimated by including among the explanatory variables following three dummy variables, Q_1, Q_2 and Q_3.

The quarterly regression model would be,

$$Y_t = \beta_0 + \beta_1 X_{1t} + \cdots + \beta_K X_{Kt} + \alpha_1 Q_{1t} + \alpha_2 Q_{2t} + \alpha_3 Q_{3t} + u_t$$

where
$$Q_{1t} = \begin{cases} 1 \text{ in the first quarter} \\ 0 \text{ in all other quarters} \end{cases}$$

$$Q_{2t} = \begin{cases} 1 \text{ in the second quarter} \\ 0 \text{ in all other quarters} \end{cases}$$

$$Q_{3t} = \begin{cases} 1 \text{ in the third quarter} \\ 0 \text{ in all other quarters} \end{cases}$$

Dummy variable trap

It should be noted that we cannot introduce here a fourth dummy variable (with values 1 for the fourth quarter and zero for all other quarters) because the determinant of the terms of sums of squares and sums of products of the explanatory variables (including the quarterly dummies) would be zero. This is due to the dummy variable X_1 which is introduced with values equal to 1 in all periods and is associated with the constant intercept β_0.

If we apply OLS to the above quarterly model the parameters estimated for the Q's will give seasonal effect for each of the three quarters. In the fourth quarter all the Q's are zero and the seasonal effect for the fourth quarter is given by the constant intercept β_0.

In fact, when we introduce a large number of dummy variables into the model, we cannot obtain the OLS estimators of the parameters. In such cases $(X'X)$ matrix may be singular and $(X'X)^{-1}$ may not exist. This problem is called **Dummy variable trap.**

Some illustrative Examples :

Example 3.12. Consider the following model showing consumption expenditure by geographical region :

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + u_i$$

where Y_i = Average consumption expenditure (₹) per person per 30 days in State i.

$$D_{1i} = \begin{cases} = 1, \text{ if the State is in the Eastern region of India} \\ = 0, \text{ otherwise (i.e., in other region of the country} \end{cases}$$

$$D_{2i} = \begin{cases} = 1, \text{ if the State is in the North-West-Central region of the countr} \\ = 0, \text{ otherwise (i.e., in other region of the country)} \end{cases}$$

Using data for 17 States of India in 2006-07 the following results are obtained by OLS method :

$$\begin{aligned} \hat{Y}_i &= 1097.38 - 241.04\, D_{1i} - 30.09\, D_{2i} \\ SE &: (103.31) \quad (133.37) \qquad (129.50) \\ t &: (10.62) \quad (-1.81) \qquad (-0.23) \end{aligned}$$

The regression results show, the mean percapita consumption expenditure in the South is about ₹ 1097.38, in the Eastern region the percapita consumption is lower by about ₹ 241.04 and that in the North-West-Central region, it is lower about ₹ 30.09.

Example 3.13. We consider a model to show the determinants of Literacy rate (in %) across 19 States of India, 2006-07.

The model takes the form :

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + u_i$$

where Y_i = literacy rate (percent)

$$D_{1i} = \text{Gender} \begin{cases} = 1 \text{ if female} \\ = 0 \text{ otherwise} \end{cases}$$

$$D_{2i} = \text{Area of residence} \begin{cases} = 1 \text{ if urban} \\ = 0 \text{ otherwise} \end{cases}$$

Using data for 19 States of India for 2006-07 the following results were obtained (by OLS method) :

$$\hat{Y}_i = 75.82 - 16.32\, D_{1i} + 16.00\, D_{2i}$$

$$\begin{array}{llll} SE : & (1.82) & (2.10) & (2.10) \\ t : & (41.65) & (-7.77) & (7.62) \end{array}$$

In this regression model there are two dummy variables. The regression results show that the mean literacy rate is about 75.82 percent. Compared with this, the average literacy rate for female is lower by about 16.32 percent, for an actual average literacy rate of (75.82 – 16.32) = 59.50 percent.

By contrast for those who live in the urban area the mean literacy rate is higher by about 16 percent, for an actual average literacy rate of (75.82 + 16) = 91.82 percent.

Example 3.14. This example shows regression with a mixture of quantitative and qualitative regressors. We consider the following model :

Let Y_i = Average consumption expenditure (₹) per person per 30 days in State i

Let, X_i = Average household size (the number of persons) in State i

$$D_{1i} = \begin{cases} = 1 \text{ if the State is in the Eastern region of India} \\ = 0 \text{ otherwise} \end{cases}$$

$$D_{2i} = \begin{cases} = 1 \text{ if the State is in the North-West-Central region of the countr} \\ = 0 \text{ otherwise} \end{cases}$$

The above equation is fitted with the help of the data on "Household consumer expenditure in India–2006-07" and obtained :

$$\hat{Y}_i = 2454.72 - 16.06\, D_{1i} + 314.02\, D_{2i} - 344.72\, X_i$$

$$\begin{array}{lllll} SE : & (505.33) & (145.22) & (165.64) & (126.49) \\ t : & (4.86) & (0.11) & (1.90) & (-2.73) \end{array}$$

$$R^2 = 0.5192$$

These results suggest that other things remaining the same, as household size goes up by one person, on an average, the percapita consumption expenditure goes down by about ₹ 344.72.

3.14. A Brief Outline on Qualitative Response Regression Models

In all the regression models that we have considered so far, we have implicitly assumed that the regressand, the dependent variable or the response variable Y is quantitative, whereas the explanatory variables are either quantitative, qualitative (or dummy), or a mixture there of.

In reality we may have to consider several models in which the regressand itself is qualitative in nature. The qualitative response regression models are now increasingly used in various areas of social sciences and medical research.

For example, we like to study the labour force participation (LFP) decision of adult males. Since an adult is either in the labour force or not, LFP is a yes or no decision. Hence the response variable or regressand, can take only two values, say, 1 if the person is in the labour force and 0 if he/she is not. In other words, the regressand is a **binary** or **dichotomous variable.**

In qualitative regression models where the regressand, Y is qualitaive and our objective is to find the probability of something happening say, $E\ (Y_i/X_{1i}, X_{2i}, ...)$. Hence, qualitative response regression models are often known as probability models.

There are four approaches to developing a probability model for a binary response variable where the regressand itself is qualitative in nature. These are :

1. The linear probability model (LPM)
2. The logit model
3. The probit model
4. The tobit model

Because of its comparative simplicity, and because it can be estimated by ordinary least square (OLS) we first consider the linear probability model (LPM).

The Linear Probability Model (LPM)

We consider a two variable regression model

$Y_i = \alpha + \beta X_i + u_i$ .... (1) where

X = family income, Y = a binary variable,

i.e., $Y = \begin{cases} = 1 \text{ if the family owns a house} \\ = 0 \text{ if it does not own a house} \end{cases}$

Model (1) looks like a typical linear regression model but because the regressand is binary it is called a **linear probability model (LPM).** This is because the conditional expectation of Y_i, given X_i, $E\left({}^{Y_i}\!/\!{}_{X_i}\right)$, can be interpreted as the conditional probability that the event will occur given X_i, that is, $P(Y_i = 1 / X_i)$. Thus, in our example $E\left({}^{Y_i}\!/\!{}_{X_i}\right)$ gives the probability of a family owning a house and whose income is the given amount X_i.

The justification of the name LPM for models like equation (1) can be seen as follows : Assuming $E\ (u_i) = 0$, as usual we obtain

$$E\left({}^{Y_i}\!/\!{}_{X_i}\right) = \alpha + \beta X_i \text{ .... (2)}$$

Now if p_i = probability that $Y_i = 1$ (that is, the event occurs), and $(1 - p_i)$ = probability that $Y_i = 0$ (that is, the event does not occur), the variable Y has the following probability distribution :

Y_i	probability
0	$1 - p_i$
1	p_i
Total	1

This shows that Y_i follows a **Bernoulli probability distribution.**

Now by definition of mathematical expectation,

we obtain : $E(Y_i) = 0 \times (1 - p_i) + 1 \times p_i = p_i$ .... (3)

Now comparing equation (2) with equation (3)

we can equate $E\left({}^{Y_i}\!/_{X_i}\right) = \alpha + \beta X_i = p_i$ .... (4)

This is in fact the conditional probability of Y_i. Since the probability p_i must lie between 0 and 1, we have the restriction,

$$0 \le E\left({}^{Y_i}\!/_{X_i}\right) \le 1 \text{ .... (5)}$$

From the above explanation it would seem that OLS can be easily extended to binary dependent variable regression models. So, we may assume that there is nothing new here. But this is not the case because the LPM poses several problems which are as follows :

(i) Non-Normality of the Disturbances u_i

Although OLS does not require the disturbances (u_i) to be normally distributed, we assumed them to be so distributed for the purpose of statistical inference. But the assumption of normality for u_i is not tenable for the LPMs because, like Y_i, the disturbances u_i, also take only two values ; that is, they also follow the Bernoulli distribution.

This can be seen clearly if we write equation (1) as $u_i = Y_i - \alpha - \beta X_i$ .... (6)

The probability distribution of u_i is :

	u_i	probability	
when $Y_i = 1$	$1 - \alpha - \beta X_i$	p_i	.... (7)
when $Y_i = 0$	$-\alpha - \beta X_i$	$1 - p_i$	

Obviously, u_i cannot be assumed to be normally distributed ; they follow the Bernoulli distribution. But the non fulfilment of the normality assumption may not be so critical as it appears because we know that the OLS point estimates still remain unbiased. Besides, as the sample size increases indefinitely, statistical theory shows that OLS estimators tend to be normally distributed generally. As a result, in large samples the statistical inference of the LPM will follow the usual OLS procedure under the normality assumption.

(ii) Heteroscedastic variances of the Disturbances

Even if $E(u_i) = 0$ and $\text{cov}(u_i, u_j) = 0$ for $i \ne j$, (i.e., no serial correlation), it can no longer be maintained that in the LPM the disturbances are homoscedastic. This is however, not surprising. We know that for a Bernoulli distribution the mean and variance are respectively p and $p\ (1-p)$ where p is the probability of success (i.e., something happening), showing that variance is a function of the mean. Hence the error variance is heteroscedastic.

In this case, var $(u_i) = p_i\ (1-p_i)$ .... (8)

Thus, the variance of the error term in the LPM is heteroscedastic.

Since $p_i = E\left({}^{Y_i}\!/_{X_i}\right) = \alpha + \beta X_i$, the variance of u_i ultimately depends on the values of X and hence is not homoscedastic.

We know that in presence of heteroscedasticity the OLS estimators are unbiased but not efficient (i.e., they do not have minimum variance). We know that there are several methods of handling the heteroscedasticity problem. Since the variance of u_i depends on $E\left({}^{Y_i}\!/_{X_i}\right)$, one way to resolve the heteroscedasticity problem is to transform the model (1) by dividing it thorough by

$$\sqrt{E\left({}^{Y_i}\!/_{X_i}\right)\left[1-E\left(\frac{Y_i}{X_i}\right)\right]} = \sqrt{p_i(1-p_i)} = \sqrt{w_i}\ \text{(say)}$$

i.e., $$\frac{Y_i}{\sqrt{w_i}} = \frac{\alpha}{\sqrt{w_i}} + \beta\frac{X_i}{\sqrt{w_i}} + \frac{u_i}{\sqrt{w_i}} \quad \text{............ (9)}$$

we can now apply OLS in this model, called **Weighted Least Square (WLS)** method with w_i serving as weights.

In theory, what we have just described is fine. But in practice the true $E\left({}^{Y_i}\!/_{X_i}\right)$ is unknown, hence the weights w_i are unknown. To estimate w_i, we can use the following two-step procedure.

Step 1 : We can run the OLS on regression (1) despite the heteroscedasticity problem and obtain

$\hat{Y}_i$ = estimate of true $E\left({}^{Y_i}\!/_{X_i}\right)$. Then obtain $\hat{w}_i = \hat{Y}_i(1-\hat{Y}_i)$, the estimate of w_i.

Step 2 : We can use the estimated w_i to transform the data shown in equation (9) and estimate the transformed equation by OLS (i.e., weighted least squares).

(iii) Non fulfilment of $0 \le E\left({}^{Y_i}\!/_{X_i}\right) \le 1$

Since $E\left({}^{Y_i}\!/_{X_i}\right)$ in the LPM measures the conditional probability of the event Y occurring, given X it must necessarily lie between 0 an 1. Although this is true *a prior*, there is no guarantee that $\hat{Y}_i$, the estimators of $E\left({}^{Y_i}\!/_{X_i}\right)$ will necessarily fulfil this restriction, and this is the real problem with the OLS estimation of the LPM. This happens because OLS does not take into account the restriction that $0 \le E(Y_i) \le 1$. There are two ways of finding out whether the estimated $\hat{Y}_i$ lie between 0 and 1. One is to estimate the LPM by the usual OLS method and find out whether the estimated $\hat{Y}_i$ lie between 0 and 1. If some are less than 0 (i.e., negative), $\hat{Y}_i$ is assumed to be zero for those cases, if they are greater than 1, they are assumed to be 1. The second procedure is to devise an estimating technique that will guarantee that the estimated conditional probabilities $\hat{Y}_i$ will lie between 0 and 1. However, in logit and probit models the estimated probabilities will indeed lie between the logical limits 0 and 1.

(iv) Questionable value of R^2 as a Measure of Goodness of Fit

The conventionally computed R^2 is of limited value in the linear probability model. Corresponding to given X, Y is either 0 or 1. Therefore all the Y values will either lie along the X axis (when $Y = 0$) or along the line corresponding to 1 (when $Y = 1$).

As a result, the conventionally computed R^2 is likely to be much lower than 1. In most practical applications the R^2 ranges between 0.2 to 0.6. However, R^2 value exceeds 0.80 when the predicted Y_i values are close to either 0 or 1.

Example 3.15. The following table (Table 3.8) gives invented data on home ownership Y (1 = owner a house, 0 = does not own a house) and family income X (thousands of dollars) for 40 families :

Table 3.8

Family	Y	X	Family	Y	X	Family	Y	X
1	0	8	15	0	6	29	0	11
2	1	16	16	1	19	30	0	10
3	1	18	17	1	16	31	1	17
4	0	11	18	0	10	32	0	13
5	0	12	19	0	8	33	1	21
6	1	19	20	1	18	34	1	20
7	1	20	21	1	22	35	0	11
8	0	13	22	1	16	36	0	8
9	0	9	23	0	12	37	1	17
10	0	10	24	0	11	38	1	16
11	1	17	25	1	16	39	0	7
12	1	18	26	0	11	40	1	17
13	0	14	27	1	20			
14	1	20	28	1	18			

From these data the estimated LPM (by OLS method) is given below :

$$\hat{Y}_i = -0.9457 + 0.1021\,X_i$$

$$SE:\ (0.1228)\qquad (0.0082)$$

$$t:\ (-7.7011)\qquad (12.4512),\ R^2 = 0.8048$$

From the estimated LPM the intercept of –0.9457 gives the 'probability' that a family with zero income will own a house. Since this value is negative and since probability cannot be negative, we take this value as zero. The slope value of 0.1021 means that for a unit change in income (here \$1000), on an average the probability of owing a house increases by 0.1021 or about 10%. Of course, given a particular level of income we can estimate the actual probability of owning a house from the fitted equation $\hat{Y}_i = -0.9457 + 0.1021\, X_i$. Thus for say $X = 12$ (\$ 12000), the estimated probability of owning a house is

$$(\hat{Y}_i / X = 12) = -0.9457 + 12 \times 0.1021 = 0.2795.$$

This is the probability that a family with an income of \$12000 will own a house is about 28 percent. We can easily estimate the probabilities $\hat{Y}_i$ for various levels of income (X_i).

If we estimate the probabilities for all given levels of income for all families, some estimated probabilities will be negative and some will exceed 1.

For instance, when $X = 8$, $(\hat{Y}_i / X = 8) = -0.9457 + 8 \times 0.1021 = -0.1289$

Similarly, when $X = 20$, $(\hat{Y}_i / X = 20) = -0.9457 + 20 \times 0.1021 = 1.0963$

This means that although $E(Y_i / X_i)$ is positive and less than 1, their estimators, $\hat{Y}_i$ need not be necessarily positive or less than 1. This is one reason that LPM is not a recommended model when the dependent variable is dichotomous. In addition to this, even if the estimated Y_i are all positive and less than 1, the LPM may suffer from the problem of heteroscedasticity. If it is so then we have to apply WLS to estimate the model.

EXERCISE

1. In a multiple linear regression model (three variable case) $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ for $1 = 1, 2, \ldots, n$, why do we insert the random disturbance term u ?
2. State the assumptions about u_i of a classical linear regression model (CLRM) where the model assumes the form $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, $i = 1, 2, \ldots, n$.
3. In a three variable linear regression model of the form $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, $i = 1, 2, \ldots, n$, how can you estimate the regression parameters by OLS method ?
4. Describe briefly the method of least squares used in estimating the regression parameters relating to a three variable linear regression model.
5. State and prove the properties of the least squares estimators relating to a three variable linear regression model (CLRM).
6. Show that in a three variable classical linear regression model, $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, $i = 1, 2, \ldots, n$, the estimated parameters coefficients are unbiased.
7. In a CLRM, $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, $i = 1, 2, \ldots, n$, express the estimated regression coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ in terms of variances and coefficient of correlations.
8. Determine the variances and covariances of the regression parameters in the model, $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, $i = 1, 2, \ldots, n$,
9. State and prove GAUSS-MARKOV THEOREM in terms of a three variable linear regression model of the form $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, $i = 1, 2, \ldots, n$.

10. What is meant by the term BLUE ? Show that in the case of a three variable linear regression model ($Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$) $\hat{\beta}$ is the BLUE of β where

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}_{3\times 1} \text{ and } \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}_{3\times 1}$$

11. How can you determine the variance of the random disturbance term u in the model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, $i = 1, 2, ..., n$?

12. In terms of a three variable linear regression model show that $\sum_{i=1}^{n} e_i^2 \Big/ (n-3)$ is an unbiased estimator of the variance of the random disturbance term σ_u^2.

13. What is a maximum likelihood estimator (MLE) ? In terms a of three variable linear regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, $i = 1, 2, ..., n$, find

(i) MLE of β where $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$, (ii) MLE of σ_u^2.

14. Show that in a three variable linear regression model MLE of $\sigma_u^2 = \sum_{i=1}^{n} e_i^2 \Big/ n$ while unbiased estimator of $\sigma_u^2 = \sum_{i=1}^{n} e_i^2 \Big/ (n-3)$

15. Show that the least square estimates of the model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ are such that :

(i) $\hat{\beta}_0 \sim N\left(\beta_0, \text{var}(\hat{\beta}_0)\right)$, (ii) $\hat{\beta}_1 \sim N\left(\beta_1, \text{var}(\hat{\beta}_1)\right)$, (iii) $\hat{\beta}_2 \sim N\left(\beta_2, \text{var}(\hat{\beta}_2)\right)$,

(iv) $u \sim N(0, \sigma_u^2)$

where $E(\hat{\beta}_0) = \beta_0$, $E(\hat{\beta}_1) = \beta_1$, $E(\hat{\beta}_2) = \beta_2$,

$$\text{var}(\hat{\beta}_0) = \frac{\sigma_u^2}{n} + \bar{X}_1^2 \, \text{var}(\hat{\beta}_1) + 2\bar{X}_1 \bar{X}_2 \, \text{cov}(\hat{\beta}_1, \hat{\beta}_2) + \bar{X}_2^2 \, \text{var}(\hat{\beta}_2),$$

$$\text{var}(\hat{\beta}_1) = \frac{\sigma_u^2}{n\sigma_{X_1}^2 \left(1 - r_{X_1X_2}^2\right)}, \quad \text{var}(\hat{\beta}_2) = \frac{\sigma_u^2}{n\sigma_{X_2}^2 \left(1 - r_{X_1X_2}^2\right)}.$$

16. Show that MLE of σ_u^2 in a three variable linear regression model is not an unbiased estimator of σ_u^2 but a consistent or asymptotically unbiased estimator of σ_u^2.

17. Discribe the testing procedure of the significance of the regression parameters of the model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ for $i = 1, 2, ..., n$.

18. What is meant by goodness of fit of a multiple linear regression model ?

19. What is meant by multiple coefficient of determination ? Derive the formula of multiple coefficient of determination, R^2 in terms of a three variable linear regression model.

20. In terms of a three variable linear regression model show that TSS = ESS + RSS where TSS = Total Sum of Squares, ESS = Explained Sum of Squares, RSS = Residual Sum of Squares.

21. In terms of a linear regression model of the form $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, $i = 1, 2, ..., n$. show that $\Sigma y_i^2 = \Sigma \hat{y}_i^2 + \Sigma e_i^2$ or, TSS = ESS + RSS

    where $\hat{y}_i = \hat{Y}_i - \bar{Y}$, $x_{1i} = X_{1i} - \bar{X}_1$, $x_{2i} = X_{2i} - \bar{X}_2$

22. In terms of a three variable linear regression model show that multiple coefficient of determination, $R^2 = \dfrac{\hat{\beta}_1 \Sigma x_{1i} y_i + \hat{\beta}_2 \Sigma x_{2i} y_i}{\Sigma y_i^2} = \dfrac{\text{ESS}}{\text{TSS}}$

    where $x_{1i} = X_{1i} - \bar{X}_1$, $x_{2i} = X_{2i} - \bar{X}_2$, $\hat{y}_i = \hat{Y}_i - \bar{Y}$

23. Define multiple coefficient of determination R^2 and adjusted $R^2(\bar{R}^2)$. Establish the relation between R^2 and $\bar{R}^2$ when there are two explanatory variables in a linear regression model.

24. What is adjusted R^2 ? Why is it used in a multiple linear regression model ?

25. Define partial correlation coefficients and the coefficient of partial determination in terms of a three variable linear regression model.

26. Establish the relation among multiple coefficient of determination R^2, the simple correlation coefficients and partial correlation coefficients in terms of a three variable linear regression model.

27. How can you formally write the regression results of the regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ where u_i (for $i = 1, 2, ..., n$) satisfies all the properties of CLRM ?

28. How can you use the analysis of variance in a three variable linear regression model ?

29. What is the meaning of the term 'prediction' ? How can you incorporate it in a three variable linear regression model ? Distinguish between point prediction and interval prediction in this regard.

30. Establish the relation between R^2 and F in terms of a three variable linear regression model. What would be the value of F when $R^2 = 0$?

31. What is an ANOVA Table ? How can you construct an ANOVA Table in terms of R^2?

32. The Cobb. Douglas production function, in its stochastic form is given by

    $Y_i = \beta_0 X_{1i}^{\beta_1} X_{2i}^{\beta_2} U_i$ where Y = output, X_1 = labour input, X_2 = capital input, U = Stochastic disturbance term, β_0 = constant technological parameter. How can you estimate the regression parameters by applying OLS method ? What stand for β_1, β_2 and $\beta_1 + \beta_2$ for the function ?

33. What do you mean by indicator variables / categorial variables / qualitative variables / dummy variables / binary variables ? Give an example.

34. What do you mean by dummy variables ? How can you incorporate these variables in the regression model. What is meant by dummy variable trap ?

35. What do you mean by dummy variables ? Explain some of the uses of dummy variables in applied economic research.

36. What are the dummy variables ? Construct a model where dummy variables are used as proxies for the dependent variable.

37. What are the dummy variables ? Construct a model where dummy variables are used as proxies of numerical factors.

38. What are the dummy variables ? Construct a model where dummy variables are used for measuring the shift of a function over time.

39. What do you mean by qualitative response regression model ? Give a brief idea of qualitative regression models widely used in applied economic research.

40. The following sums were obtained from 10 sets of observations on Y, X_1 and X_2 : $\Sigma Y = 20$, $\Sigma X_1 = 30$, $\Sigma X_2 = 40$, $\Sigma Y^2 = 88.2$, $\Sigma X_1^2 = 92$, $\Sigma X_2^2 = 163$, $\Sigma YX_1 = 59$, $\Sigma YX_2 = 88$, $\Sigma X_1X_2 = 199$. Estimate the regression of Y on X_1 and X_2 and test the hypothesis that the coefficient of X_2 is zero.

41. Consider the following regression model in deviation form : $y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$ with sample data : $n = 100$, $\Sigma y^2 = \frac{493}{3}$, $\Sigma x_1^2 = 30$, $\Sigma x_2^2 = 3$, $\Sigma x_1 y = 30$, $\Sigma x_2 y = 20$, $\Sigma x_1 x_2 = 0$

(i) Compute OLS estimates of β_1, β_2 and R^2

(ii) Test the hypothesis $H_0 : \beta_2 = 7$ against $\beta_2 \neq 7$

(iii) Test the hypothesis $H_0 : \beta_1 = \beta_2 = 0$ against $H_1 : \beta_1 \neq 0$, or, $\beta_2 \neq 0$.

(iv) Test the hypothesis $H_0 : \beta_2 = 7\,\beta_1$ against $H_1 : \beta_2 \neq 7\beta_1$.

42. A production function model is specified as $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ where $Y_i =$ log output, X_{1i} log labour input, and $X_{2i} =$ log capital input. The data refer to a sample of 23 firms and observations are measured as devivations from the sample means :

$\Sigma x_{1i}^2 = 12$, $\Sigma x_{1i} x_{2i} = 8$, $\Sigma x_{2i}^2 = 12$, $\Sigma x_{1i} y_i = 10$, $\Sigma x_{2i} y_i = 8$, $\Sigma y_i^2 = 10$

(i) Estimate β_1, β_2 and their standard errors.

(ii) Find R^2 and adjusted R^2.

(iii) Test the hypothesis that $\beta_1 + \beta_2 = 1$

(iv) Suppose now that you wish to impose the *apriori* restriction that $\beta_1 + \beta_2 = 1$. What is the least squares estimate of β_1 and its standard error ? What is the value of R^2 in this case ? Compare these results with those obtained in (i) and comment.

43. The following table shows 10 sets of values of three variables Y (dependent variable), X_1 and X_2 (two independent variables).

Y :	3.5	4.3	5	6	7	9	8	10	12	14
X_1 :	15	20	30	42	50	54	65	72	85	90
X_2 :	16	13	10	7	7	5	4	3	3.5	2

(i) Consider a model of the form $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$. Find the least squares regression equation of Y on X_1 and X_2.

(ii) Compute the coefficient of multiple determination and the standard errors of the estimated parameters and conduct tests of significance.

(iii) Construct 95 percent confidence intervals for the population parameters and σ_u^2.

(iv) Find the explained and unexplained variation in Y.

44. The following results were obtained from a sample of 12 firms on their output (Y), labour input (X_1) and capital input (X_2), measured in arbitrary units :

$\Sigma Y = 753$, $\Sigma Y^2 = 48{,}139$, $\Sigma X_1 Y = 40{,}830$, $\Sigma X_1 = 643$, $\Sigma X_1^2 = 34{,}834$, $\Sigma X_2 Y = 6{,}796$,

$\Sigma X_2 = 106$, $\Sigma X_2^2 = 976$, $\Sigma X_1 X_2 = 5{,}779$.

(i) Find the least squares equation of Y on X_1 and X_2. What is the economic meaning of your coefficients ?

(ii) Given the following sample values of output (Y), compute the standard errors of the estimates and test their statistical significance.

Firms :	A	B	C	D	E	F	G	H	I	J	K	L
Output :	64	71	53	67	55	58	77	57	56	51	76	68

(iii) Find the multiple correlation coefficient and the unexplained variation in ouput.

(iv) Constant 99% confidence intervals for the population parameters.

45. The following table shows the value of imports (Y), the level of Gross National Product (X_1) measured in arbitrary units and the price index of imported goods (X_2), over the twelve-year period for a certain country.

	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Y:	57	43	73	37	64	48	56	50	39	43	69	60
X_1:	220	215	250	241	305	258	354	321	370	375	385	385
X_2:	125	147	118	160	128	149	145	150	140	115	155	152

(i) Estimate the import function $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$.

(ii) What is the economic meaning of your estimates ?

(iii) Construct tests of significance for the regression estimates at 5% and 1% levels of significance.

(iv) Compute R^2 and adjusted R^2.

46. The following table includes the output (Y), the labour input (L) and capital input (K) of 15 firms of the chemical industry :

Firms :	1	2	3	4	5	6	7	8
Y (1000 tons) :	60	120	190	250	300	360	380	430
L (hours) :	1100	1200	1430	1500	1520	1620	1800	1820
K (machine hours) :	300	400	420	400	510	590	600	630
	9	10	11	12	13	14	15	
	440	490	500	520	540	410	350	
	1800	1750	1950	1960	1830	1900	1500	
	610	630	850	900	980	900	800	

(i) Fit a Cobb-Douglas production function to the above data $Y = \beta_0 L^{\beta_1} K^{\beta_2} u$.

(ii) Construct appropriate tests of significance of the parameter estimates at 5% and 1% levels of significance.

(iii) What are the margnal and average productivities of the factors L and K ?

(iv) What do your results suggest regarding the returns to scale ?

47. The following table shows the price index of durables, the average yearly income and expenditure on durables of a 'typical' household of a country.

Year :	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Expenditure on durables (Y) in £ :	115	110	115	120	140	100	105	95	135	105
Household income (X_1) £ :	1855	2000	2010	2040	2275	2255	1995	1905	2355	2035
Price Index (X_2) :	100	102	95	95	94	110	110	112	115	120

(i) Fit a regression line to the function : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$

(ii) Test your results by using the Analysis of variance table.

48. The following table shows the consumption of tobacco manufactures, consumer's income and the price of tabacco manufactures for France during 1950s.

Year	Consumption (million tons) D	Income (million francs) Y	Price of tobacco (francs per kg) P
1950	59,190	76,200	23.56
1951	65,450	91,700	24.44
1952	62,360	106,700	32.07
1953	64,700	111,600	32.46
1954	67,400	119,000	31.15
1955	64,440	129,200	34.14
1956	68,000	143,400	35.30
1957	72,400	159,600	38.70
1958	75,710	180,000	39.63
1959	70,680	193,000	46.68

(i) Fit a linear regression $D = \beta_0 + \beta_1 P + \beta_2 Y + u$ and a non linear function of the constant elasticity type $D = \beta_0 P^{\beta_1} Y^{\beta_2} u$

(ii) Conduct tests of significance using the analysis of variance table.

(iii) Compare the price and income elasticities of the two functions.

49. In a multiple regression equation $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$ explain how you will test the joint hypothesis $\beta_1 = \beta_2$ and $\beta_2 = 1$.

50. Consider the following regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ where $u_i \sim N(0, \sigma_u^2)$.

The following data set are given below :

Y :	4	7	3	9	7
X_1 :	2	3	1	5	9
X_2 :	5	3	2	1	7

(i) Estimate $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$.

(ii) Find out $\text{var}(\hat{\beta}_0)$, $\text{var}(\hat{\beta}_1)$, $\text{var}(\hat{\beta}_2)$ and $\text{cov}(\hat{\beta}_1, \hat{\beta}_2)$.

(iii) Find R^2 and adjusted R^2.

(iv) Write the regression results in the summary form.

(v) Test $H_0 : \beta_0 = 0$ against $H_1 : \beta_0 \neq 0$ and find 99% and 95% confidence intervals of β_0.

(vi) Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ and find 99% and 95% confidence intervals of β_1.

(vii) Test $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$ and find 99% and 95% confidence intervals of β_2.

(viii) Test $H_0 : \beta_1 = \beta_2$ against $H_1 : \beta_1 \neq \beta_2$ and find 99% and 95% confidence internals of $(\beta_1 - \beta_2)$.

(ix) Construct 99% and 95% confidence intervals of σ_u^2.

(x) Find out the point predictor of Y when $X_1 = 10$ and $X_2 = 8$.

(xi) Construct 95% confidence interval for the prediction.

(xii) Test whether the effect of X_2 on Y is more important than the effect of X_1 on Y

51. Eight students made the following scores on pretest, test and final examinations in a certain subject. Fit the linear regression equation of final exmination score (Y) on pretest score (X_1) and test score (X_2) :

Students :	1	2	3	4	5	6	7	8
Pretest score (X_1) :	43	38	27	28	35	21	19	13
Test score (X_2) :	22	29	23	33	20	8	17	19
Final score (Y) :	66	38	55	63	25	17	33	18

Assume a linear regression equation of the form :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i,\ u_i \sim N(0, \sigma_u^2),\ i = 1, 2, ..., 8$$

(i) Find $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$

(ii) Find $\text{var}(\hat{\beta}_0)$, $\text{var}(\hat{\beta}_1)$, $\text{var}(\hat{\beta}_2)$ and $\text{cov}(\hat{\beta}_1, \hat{\beta}_2)$

(iii) Find R^2 and adjusted R^2

(iv) Write the regression results in the summary form.

(v) Test whether $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ are significant or not at 5% level of significance.

(vi) Find out the point predictor of Y when $X_1 = 44$ and $X_2 = 62$

(vii) Construct 95% confidence interval of σ_u^2.

(viii) Construct 95% confidence interval of the prediction.

52. Following exercise–50, test the regression results by using Analysis of variance table.

53. Following exercise–51, test the regression results by using Analysis of variance table.

54. The following table shows monthly income (₹) (Y), monthly savings (₹) (X_1) and age (X_2) of 10 persons.

Y:	2000	8000	10,000	12,000	20,000	25,000	28,000	30,000	36,000	40,000
X_1:	10,000	20,000	25,000	30,000	40,000	45,000	50,000	55,000	60,000	70,000
X_2:	22	25	30	35	41	43	46	50	57	60

Assume a linear regression equation of the form : $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$

and assume X_2 as a dummy variable, $X_2 = \begin{cases} = 0 \text{ if} <40 \text{ years of age} \\ = 1 \text{ if} \geq 40 \text{ years of age} \end{cases}$

(i) Find $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$.

(ii) Find R^2 and adjusted R^2.

(iii) Write the regression results in the summary form.

(iv) Test whether $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ are significant or not at 5% level of significance.

(v) Construct 95% confidence interval of σ_u^2.

55. The following regression was estimated from 16 quarterly observations (t ratios in parentheses) :

$$Y_t = 70.7 - 0.90\, X_t + 0.43\, S_{1t} + 6.55\, S_{2t} - 2.83\, S_{3t},\ R^2 = 0.68$$
$$(3.7)\quad (0.27)\qquad (3.37)\qquad (3.40)\qquad (3.37)$$

where $S_{it} = 1$ in the ith quarter and 0 otherwise. Explain the implied pattern of seasonal variation and interpret the result.

56. You are given the following regression results :

$$\hat{Y}_t = 16.899 - 2972.5 X_{1t},\ R^2 = 0.6149$$
$$t : (8.5152)\quad (-4.7280)$$

$$\hat{Y}_t = 9734.2 - 3782.2\, X_{1t} + 2815\, X_{2t},\ R^2 = 0.7706$$
$$t : (3.3705)\quad (-6.6070)\quad (2.9712)$$

Can you find out the sample size underlying these results ?

[**Hints** : Use the relationship among R^2, F and t values.]

57. From the data for 46 States in the Unites States for a given year the following regression results were obtained :

$$\log \hat{C} = 4.30 - 1.34 \log P + 0.17 \log Y$$
$$SE \quad : (0.91)\quad (0.32)\qquad (0.20)\qquad \bar{R}^2 = 0.27$$

where C = units of consumption of a commodity per year

P = real price per unit of the commodity

Y = percapita real disposable income.

(i) What is the elasticity of demand for the commodity with respect to price ? Is it statistically significant ? If so, is it statistically different from 1 ?

(ii) What is the income elasticity of demand for the commodity ? Is it statistically significant ?

(iii) How would you retrieve R^2 from $\bar{R}^2$ given above ?

58. From a sample of 209 firms the following regression results were obtained :

$$\text{Log (salary)} = 4.32 + 0.280 \log (\text{sales}) + 0.0174\ \text{roe} + 0.00024\ \text{ros}\quad R^2 = 0.283$$
$$SE \quad : (0.32)\quad (0.035)\qquad (0.0041)\qquad (0.00054)$$

where salary = salary of CEO
sales = annual firm sales
roe = return on equity in percent
ros = return on firm's stock.

(i) Interpret the regression taking into account any prior expectations that you may have about the signs of the various coefficients.

(ii) Which of the coefficients are individually statistically significant at the 5% level?

(iii) What is the over all significance of the regression ?

(iv) Can you interpret the coefficients of roe and ros as elasticity coefficients ? Why or why not ?

59. Consider the following wage-determination equation for the British economy for the period 1950-1969.

$$\hat{W}_t = 8.582 + 0.364(PF)_t + 0.004(PF)_{t-1} - 2.560U_t \quad R^2 = 0.873, \text{ df} = 15$$

$$SE : (1.129) \quad (0.080) \quad (0.72) \quad (0.658)$$

where W = wages and salaries per employee
PF = price of final output at factor cost
U = unemployment in Great Britain as a percentage of the total number of employees in Great Britain
t = time.

(i) Interpret the regression equation.

(ii) Are the estimated coefficients individually significant ?

(iii) What is the rationale for the introduction of $(PF)_{t-1}$?

(iv) How would you compute elasticity of wages and salaries per employee with respect to unemployment rate U ?

60. Consider the following data set :

Y :	1	3	8
X_1 :	1	2	3
X_2 :	2	1	-3

Based on these data, estimate the following regressions :

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + u_{1i} \quad \text{..............} (1)$$

$$Y_i = \lambda_0 + \lambda_2 X_{2i} + u_{2i} \text{..............} (2)$$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \quad \text{.....}(3)$$

(i) Estimate the regression coefficients in each case.

(ii) Is $\alpha_1 = \beta_1$? Why or why not ?

(iii) Is $\lambda_2 = \beta_2$? Why or why not ?

What important conclusion do you draw from this exercise ?

61. From the following data estimate the partial regression coefficients, their standard errors and the adjusted and unadjusted R^2 values :

$\bar{Y} = 367.693$, $\bar{X}_1 = 402.760$, $\bar{X}_2 = 8.0$

$\Sigma(Y_i - \bar{Y})^2 = 66042.269$, $\Sigma(X_{1i} - \bar{X}_1)^2 = 84855.096$

$\Sigma(X_{2i}-\bar{X}_2)^2 = 280.000$, $\Sigma(Y_i-\bar{Y})(X_{1i}-\bar{X}_1) = 74778.346$

$\Sigma(Y_i-\bar{Y})(X_{2i}-\bar{X}_2) = 4250.900$, $\Sigma(X_{1i}-\bar{X}_1)(X_{2i}-\bar{X}_2) = 4796.00$, $n = 15$

62. Is it possible to obtain the following from a set of data ?

(i) $r_{23} = 0.9$, $r_{12} = -0.2$, $r_{12} = 0.8$

(ii) $r_{12} = 0.6$, $r_{23} = -0.9$, $r_{31} = -0.5$

(iii) $r_{21} = 0.01$, $r_{13} = 0.66$, $r_{23} = -0.7$

63. The regressed child mortality (*CM*) on percapita GNP (PGNP) and the female literacy rate (FLR) for a sample of 64 countries is given below :

$$\widehat{CM}_i = 263.6416 - 0.0056\ PGNP_i - 2.2316\ FLR_i$$

$$SE : (11.5932) \qquad (0.0019) \qquad (0.2099)$$

$$R^2 = 0.7077, \ \bar{R}^2 = 0.6981$$

(i) Interpret the regression results.

(ii) What about the statistical significance of the observed results ?

(iii) Is the coefficient of PGNP of – 0.0056 statistically significant ?

(iv) Is the coefficient of FLR of – 2.2316 statistically significant ?

(v) Are both the coefficients statistically significant jointly ?

4

Violations of Classical Assumptions—The Problems of Heteroscedasticity, Autocorrelation and Multicollinearity

4.1. Introduction

Let us consider a two-variable linear regression model $Y_i = \alpha + \beta X_i + u_i$ for $i = 1, 2, \ldots\ldots, n$.

If the model is a classical linear regression model (CLRM) then we have the following assumptions :

(*i*) u_i is a random variable

(*ii*) $E(u_i) = 0$ for each i

(*iii*) $E(u_i^2) = \sigma_u^2$ or σ^2 (constant)

(*iv*) $Cov\ (u_i, u_j) = E(u_i, u_j) = 0$ for $i \neq j$

(*v*) X is non-stochastic or non-random.

We now put special consideration of assumptions (*iii*) and (*iv*). Assumption (*iii*) $E(u_i^2) = \sigma^2$ means that the variance of the disturbance term is constant. In other words the probability distribution of the disturbance term does not vary from each other. This feature of homogeneity of variance (or constant variance) is known as *homoscedasticity*. It may be the case, however, that all of the disturbance terms do not have the same variance. This condition of non-constant variance or non-homogeneity of variance is known as *heteroscedasticity*. Thus, we say that u_i's are heteroscedastic when $\text{var}\ (u_i) \neq \sigma_u^2$ (or σ^2) [a constant value] but $\text{var}\ (u_i) = E(u_i^2) = \sigma_{u_i}^2$ (a value that varies).

Also we should not assume that each disturbance term has the same expected value equal to zero (*i.e.*, $E(u_i) = 0$.

If all the disturbance terms have expected value zero and same variance σ^2, then we say that all the disturbance terms are identically distributed.

If (*iv*) is also satisfied, it means that the different disturbance terms are independent of each other. So, when all (*ii*), (*iii*) and (*iv*) are satisfied then we can say that the different disturbance terms are identically and independently distributed.

If the disturbance term varies from observation to observation, the different disturbance terms are not identically distributed.

Here $\text{var}(u_i) = E(u_i^2) = \sigma_{u_i}^2 \neq \sigma_{u_j}^2$ when $i \neq j$.

This is the problem of *Heteroscedasticity*. The CLRM assumes that variance of the disturbance term is constant and if it is not constant, then the problem of heteroscedasticity is said to exist in the CLRM. But in the presence of the problem of heteroscedasticity the disturbance terms are independent of each other. Here $cov(u_i, u_j) = 0$, so assumption (*iv*) of CLRM remains valid. So, if there is the problem of heteroscedasticity only assumption (*iii*) of CLRM is violated.

There is another problem, termed as the problem of Autocorrelation.

Assumption (*iv*) of CLRM is given by,

$$Cov(u_i, u_j) = E\left[\{u_i - E(u_i)\}\{u_j - E(u_j)\}\right]$$

$$= E\,(u_i, u_j) = E(u_i).E(u_j) = 0 \text{ (for } i \neq j)$$

$$\therefore E(u_i) = E(u_j) = 0$$

This assumption implies that the successive values of disturbance term u are temporarily independent, i.e., disturbance occurring at one point of observation is not related to any other disturbance. In other words, when observations are taken over time, the effect of disturbance occurring at one period does not carry over into another period.

If the above assumption is not satisfied i.e., if the value of u in any particular period is correlated with its own preceding value(s), we say that there is *autocorrelation* or *serial correlation* of the random variables.

Autocorrelation is a special case of correlation. It refers to the relationship between the successive values of the same variable in different periods, while correlation refers to the relationship between two or more different variables. We can readily observe autocorrelation in the time series sample which exhibits a secular trend or long run movement over time. Cyclical fluctuations also impose regularity among successive observations of the variables over time and, therefore, are the cause of autocorrelation. It should be noted that autocorrelation does not arise in cross section data. For example, in a cross section sample, giving data on income and expenditure of different families, the dependence between the expenditure behaviours of two families is completely ruled out.

Thus, in the presence of autocorrelation the different disturbance terms are not independent of each other. This means that cov $(u_i, u_j) = E\,(u_i, u_j) \neq 0$. But here assumption (iii) remains valid when the disturbance terms are identically distributed but not independently distributed.

So, both the assumptions (iii) and (iv) of CLRM are violated if there is the problem of heteroscedasticity as well as the problem of autocorrelation.

4.2. Matrix Representation of Autocorrelation and Heteroscedasticity

Let us consider the dispersion matrix or variance-covariance matrix of the dispersion vector

$$D(u) = E(uu') \text{ where } u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}_{n\times 1} \text{ and } u' = [u_1, u_2, ... u_n]_{1 \times n}$$

$$\left.\begin{array}{l} \text{Since } Y_i = \alpha + \beta X_i + u_i \text{ for } i = 1, 2, ..., n \\ Y_1 = \alpha + \beta X_1 + u_1 \\ Y_2 = \alpha + \beta X_2 + u_2 \\ \text{..................................} \\ Y_n = \alpha + \beta X_n + u_n \end{array}\right\}$$

$$\text{Now, } D(u) = E(uu') = E\begin{bmatrix} u_1 \\ u_2 \\ \cdots \\ u_n \end{bmatrix}[u_1, u_2, ..., u_n]$$

$$= E\begin{bmatrix} u_1^2 & u_1u_2 & \cdots & u_1u_n \\ u_2u_1 & u_2^2 & \cdots & u_2u_n \\ \cdots & \cdots & \cdots & \cdots \\ u_nu_1 & u_nu_2 & \cdots & u_n^2 \end{bmatrix}_{n\times n} = \begin{bmatrix} E(u_1^2) & E(u_1u_2) & \cdots & E(u_1u_n) \\ E(u_2u_1) & E(u_2^2) & \cdots & E(u_2u_n) \\ \cdots & \cdots & \cdots & \cdots \\ E(u_nu_1) & E(u_nu_2) & \cdots & E(u_n^2) \end{bmatrix}_{n\times n}$$

If the model is CLRM, then by assumption $E(u_i^2) = \sigma_u^2$ (constant) and $E(u_i, u_j) = 0$ for $i \neq j$.

In that case

$$D(u) = \begin{bmatrix} \sigma_u^2 & 0 & \cdots & 0 \\ 0 & \sigma_u^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sigma_u^2 \end{bmatrix} = \sigma_u^2.I_n$$

$$\text{where } I_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \text{ is an identity matrix.}$$

But if there is the problem of heteroscedasticity and/or autocorrelation, then $D(u) \neq \sigma_u^2.I_n$.

If there is the problem of heteroscedasticity, then $E(u_i^2) = \sigma_{u_i}^2$ and $E(u_i, u_j) = 0$ for $i \neq j$. In that case $D(u)$ has non-zero diagonal terms but zero non-diagonal terms. However, the diagonal terms are different. So, the dispersion matrix of the disturbance vector $D(u)$ is a diagonal matrix but not a scalar matrix.

$$\text{i.e., } D(u) = \begin{bmatrix} \sigma_{u_1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{u_2}^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sigma_{u_n}^2 \end{bmatrix}$$

If there is the problem of autocorrelation, then $E(u_i^2)=\sigma_u^2$ but $E(u_i, u_j)\neq 0$. Hence the non-diagonal terms of the dispersion matrix $D(u)$ are not zero. This means that the dispersion matrix $D(u)$ is not a diagonal matrix. In that case we can write $D(u)=E(uu')$ $=\sigma_u^2\Omega$ where $\Omega\neq I_n$ where I_n is an identity matrix.

Following three examples of dispersion matrices help to understand the concept of autocorrelation and heteroscedasticity.

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix} \qquad \begin{bmatrix} 1 & \frac{1}{2} & 0 \\ \frac{1}{2} & 1 & \frac{1}{2} \\ 0 & \frac{1}{2} & 1 \end{bmatrix} \qquad \begin{bmatrix} 2 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 3 \end{bmatrix}$$

Heteroscedasticity with no autocorrelation Homoscedasticity with autocorrelation Heteroscedasticity with autocorrelation

4.3. Consequences of the Problems of Autocorrelation and Heteroscedasticity

If there is the problem of heteroscedasticity or autocorrelation, applying the OLS method we can get the unbiased estimators of the parameters but cannot get the BLUE of the parameters.

4.4. Consequences of the Problem of Heteroscedasticity

In the presence of the problem of heteroscedasticity the OLS estimators will be unbiased but we cannot get the BLUE of the parameters. This proposition can be examined as follows. Let us consider a simple model $y_i=\beta x_i+u_i$ for $i=1, 2, \ldots n$, with $E(u_i)=0$, $E(u_i, u_j)=0$ for $i\neq j$, $E(u_i^2)=\sigma_{u_i}^2$ where $y_i=Y_i-\overline{Y}$ and $x_i=X_i-\overline{X}$.

[Since $Y_i=\alpha+\beta X_i+u_i$ $\therefore \overline{Y}=\alpha+\beta\overline{X}+\overline{u}$, and $\overline{u}=0$.

Now, $Y_i-\overline{Y}=\beta(X_i-\overline{X})+u_i$ or, $y_i=\beta x_i+u_i$]

The OLS estimator of β is denoted by $\hat{\beta}=\dfrac{\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}x_i^2}=\dfrac{\sum_{i=1}^{n}x_i(\beta x_i+u_i)}{\sum_{i=1}^{n}x_i^2}$

$$=\beta\frac{\sum_{i=1}^{n}x_i^2}{\sum_{i=1}^{n}x_i^2}+\frac{\sum_{i=1}^{n}x_iu_i}{\sum_{i=1}^{n}x_i^2}$$

$$\text{or, } \hat{\beta}=\beta+\frac{\sum_{i=1}^{n}x_iu_i}{\sum_{i=1}^{n}x_i^2} \quad \text{or, } E(\hat{\beta})=E(\beta)+\frac{\sum_{i=1}^{n}x_iE(u_i)}{\sum_{i=1}^{n}x_i^2}$$

or, $E(\hat{\beta})=\beta+0=\beta$, $[\because E(u_i)=0]$ $\therefore E(\hat{\beta})=\beta$

Similarly, $\hat{\alpha}=\overline{Y}-\hat{\beta}\overline{X}=(\alpha+\beta\overline{X}+\overline{u})-\hat{\beta}\overline{X}$

$$\therefore E(\hat{\alpha})=\alpha+\beta\overline{X}+E(\overline{u})-E(\hat{\beta}).\overline{X}$$

$$=\alpha+\beta\overline{X}+E(\overline{u})-\beta\overline{X} \qquad \because \overline{u}=0,\text{ and } E(\overline{u})=0$$

$\therefore E(\hat{\alpha})=\alpha$

This proves that the least square estimators are unbiased even under the condition of heteroscedasticity.

Now, we have to show that the variance of $\hat{\beta}$ is not minimum.

$$\text{Since } \hat{\beta}=\beta+\frac{\sum_{i=1}^{n}x_iu_i}{\sum_{i=1}^{n}x_i^2} \quad \text{or, } (\hat{\beta}-\beta)=\sum_{i=1}^{n}x_iu_i\Big/\sum_{i=1}^{n}x_i^2$$

$$\text{Now, } \operatorname{var}(\hat{\beta})=E\left[\hat{\beta}-E(\hat{\beta})\right]^2=E\left[\hat{\beta}-\beta\right]^2=E\left[\frac{\sum_{i=1}^{n}x_iu_i}{\sum_{i=1}^{n}x_i^2}\right]^2$$

$$\therefore \operatorname{var}(\hat{\beta})=E\left[\frac{\sum_{i=1}^{n}x_iu_i}{\sum_{i=1}^{n}x_i^2}\right]^2$$

$$=E\left[\frac{\sum_{i=1}^{n}x_i^2u_i^2+2\sum_i\sum_j x_ix_ju_iu_j}{\left(\sum_{i=1}^{n}x_i^2\right)^2}\right]=\frac{\sum_{i=1}^{n}x_i^2E(u_i^2)}{\left(\sum_{i=1}^{n}x_i^2\right)^2}+\frac{2\sum_i\sum_j x_ix_jE(u_iu_j)}{\left(\sum_{i=1}^{n}x_i^2\right)^2}$$

$$=\frac{\sum_{i=1}^{n}x_i^2\sigma_{u_i}^2}{\left(\sum_{i=1}^{n}x_i^2\right)^2}+0 \quad \left[\begin{array}{l}\because E(u_i^2)=\sigma_{u_i}^2\text{ (not constant)}\\ E\, u(u_j =)0 \text{ for } i\neq j\end{array}\right] \quad \therefore \operatorname{var}(\hat{\beta})=\frac{\sum_{i=1}^{n}x_i^2\sigma_{u_i}^2}{\left(\sum_{i=1}^{n}x_i^2\right)^2}$$

From the minimum variance property of the OLS estimator we know that

$\text{var}(\hat{\beta}) = \sigma_u^2 \sum_{i=1}^{n} k_i^2 = \dfrac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2}$. But under heteroscedastic assumption we get,

$$\text{var}(\hat{\beta}) = \sum_{i=1}^{n} k_i^2 E(u_i^2) = \sum_{i=1}^{n} k_i^2 \sigma_{u_i}^2 \neq \sigma_u^2 \sum_{i=1}^{n} k_i^2 \quad \text{or,} \quad \left(\neq \frac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2} \right).$$

Since $\text{var}(\hat{\beta}) = \dfrac{\sigma_u^2}{\sum_{i=1}^{n} x_i^2}$ when there is no problem of heteroscedasticity.

So, $\text{var}(\hat{\beta})$ in the presence of the problem of heteroscedasticity is not equal to $\text{var}(\hat{\beta})$ in the CLRM.

Hence, BLUE may not be satisfied. Furthermore, in the presence of heteroscedasticity the OLS estimators will be inefficient and tests of significance will not be applicable.

4.5. Method for Estimating Regression Parameters in the Presence of the Problem of Heteroscedasticity

We have to find out the appropriate method for estimating the parameters of the model in the presence of the problem of heteroscedasticity.

Let us consider the following model :

$y_i = \beta x_i + u_i$ (where $Y_i = \alpha + \beta X_i + u_i$, $y_i = Y_i - \bar{Y}, x_i = X_i - \bar{X}, \bar{u} = 0$)

In the presence of the problem of heteroscedasticity **we assume that** $E(u_i^2) = \sigma_u^2 x_i$

This means that the variance of the disturbance terms varies positively with the independent variable (x_i).

If we apply the OLS method, then we have, $\hat{\beta} = \beta + \dfrac{\sum_{i}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2}$.

Now, $E(\hat{\beta}) = \beta$ because $E(u_i) = 0$.

and $\text{var}(\hat{\beta}) = E[\hat{\beta} - E(\hat{\beta})]^2$

$$= E\left[\frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2}\right]^2 = E\left[\frac{\sum_{i=1}^{n} x_i^2 u_i^2}{\left(\sum_{i=1}^{n} x_i^2\right)^2}\right] \quad \text{or,} \quad \text{var}(\hat{\beta}) = \frac{\sum_{i=1}^{n} x_i^2 E(u_i^2)}{\left(\sum_{i=1}^{n} x_i^2\right)^2} = \frac{\sum_{i=1}^{n} x_i^2 \sigma_u^2 x_i}{\left(\sum_{i=1}^{n} x_i^2\right)^2}$$

$$\therefore \text{var}(\hat{\beta}) = \frac{\sigma_u^2 \sum_{i=1}^{n} x_i^3}{\left(\sum_{i=1}^{n} x_i^2\right)^2} \quad [\because E(u_i^2) = \sigma_u^2 x_i]$$

We now consider an *alternative method* of estimating β which is called **weighted least squares (WLS) method [Model (2)].**

Dividing both sides of the equation $y_i = \beta x_i + u_i$ ...... (1) by $\sqrt{x_i}$ we get,

$$\frac{y_i}{\sqrt{x_i}} = \beta\sqrt{x_i} + \frac{u_i}{\sqrt{x_i}}, \quad \text{.........(2)}$$

$$\text{Now, } E\left(\frac{u_i}{\sqrt{x_i}}\right) = \frac{E(u_i)}{\sqrt{x_i}} = 0 \qquad [\because E(u_i) = 0]$$

$$\text{and } \text{var}\left(\frac{u_i}{\sqrt{x_i}}\right) = E\left[\frac{u_i}{\sqrt{x_i}}\right]^2 = \frac{1}{x_i} \cdot \sigma_u^2 \cdot x_i = \sigma_u^2$$

$$\therefore \text{var}\left(\frac{u_i}{\sqrt{x_i}}\right) = \sigma_u^2.$$

So, variance of the disturbance term of model (2) is constant and hence model (2) is free from the problem of heteroscedasticity. It satisfies the CLRM properties.

Since the new model (2) is equivalent to a CLRM we can apply OLS method to this model (2). The OLS method applied to the transformed equation (2) is called WLS method.

Let us suppose that β^* is the WLS estimator of β. Then,

$$\beta^* = \frac{\sum_{i=1}^{n} \frac{y_i}{\sqrt{x_i}} \cdot \sqrt{x_i}}{\sum_{i=1}^{n} (\sqrt{x_i})^2} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\bar{y}}{\bar{x}} \quad \therefore \beta^* = \sum_{i=1}^{n} y_i \Big/ \sum_{i=1}^{n} x_i = \frac{\bar{y}}{\bar{x}}$$

$$\text{Now, } \beta^* = \sum_{i=1}^{n} y_i \Big/ \sum_{i=1}^{n} x_i = \frac{\sum_{i=1}^{n} (\beta x_i + u_i)}{\sum_{i=1}^{n} x_i} = \beta\frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i} + \frac{\sum_{i=1}^{n} u_i}{\sum_{i=1}^{n} x_i} = \beta + \frac{\sum_{i=1}^{n} u_i}{\sum_{i=1}^{n} x_i}$$

$$\therefore \beta^* = \beta + \left(\sum_{i=1}^{n} u_i \Big/ \sum_{i=1}^{n} x_i\right)$$

$$\text{Now, } E(\beta^*) = E(\beta) + \sum_{i=1}^{n} E(u_i) \Big/ \sum_{i=1}^{n} x_i$$

$= \beta + 0$, $[\because E(u_i) = 0]$

$= \beta$

$\therefore E(\beta^*) = \beta.$

Now, $\text{var}(\beta^*) = E[\beta^* - E(\beta^*)]^2$

$= E[\beta^* - \beta]^2 \quad [\because E(\beta^*) = \beta]$

$$= E\left[\frac{\sum_{i=1}^{n} u_i}{\sum_{i=1}^{n} x_i}\right]^2 \qquad \left[\because \beta^* = \beta + \left(\sum_{i=1}^{n} u_i \Big/ \sum_{i=1}^{n} x_i\right) \text{ or, } \beta^* - \beta = \sum_{i=1}^{n} u_i \Big/ \sum_{i=1}^{n} x_i\right]$$

$$= \frac{\sum_{i=1}^{n} E(u_i^2)}{\left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n} \sigma_u^2 x_i}{\left(\sum_{i=1}^{n} x_i\right)^2} \quad \left[\because E(u_i^2) = \sigma_u^2 x_i\right]$$

$$\text{or, } \text{var}(\beta^*) = \frac{\sigma_u^2 \sum_{i=1}^{n} x_i}{\left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sigma_u^2}{\sum_{i=1}^{n} x_i} \quad \text{But, } \text{var}(\hat{\beta}) = \frac{\sigma_u^2 \sum_{i=1}^{n} x_i^3}{\left(\sum_{i=1}^{n} x_i^2\right)^2}$$

$$\text{As } \left(\sum_{i=1}^{n} x_i^3\right)\sum_{i=1}^{n} x_i > \left(\sum_{i=1}^{n} x_i^2\right)^2 \text{ or, } \frac{\sum_{i=1}^{n} x_i^3}{\left(\sum_{i=1}^{n} x_i^2\right)^2} > \frac{1}{\sum_{i=1}^{n} x_i} \text{ or, } \frac{\sigma_u^2 \sum_{i=1}^{n} x_i^3}{\left(\sum_{i=1}^{n} x_i^2\right)^2} > \frac{\sigma_u^2}{\sum_{i=1}^{n} x_i}$$

$\therefore \text{var}(\hat{\beta}) > \text{var}(\beta^*)$

So, WLS estimator of the parameter β, denoted by β^* has a lower variance than the variance of $\hat{\beta}$.

In this sense WLS method is more appropriate to a model which is subject to the problem of heteroscedasticity.

[**Note** : Heteroscedasticity may also be of the form $\sigma_{u_i}^2 = k^2 x_i^2$ i.e., $E(u_i^2) = k^2 x_i^2$]

4.6. Tests for Heteroscedasticity

There are three important tests for heteroscedasticity.

(*i*) Spearman's Rank correlation test.

(*ii*) Goldfeld and Quandt test.

(*iii*) Glejser's test.

All these test criteria are based on the OLS residual terms.

Let us consider the following model : $Y_i = \alpha + \beta X_i + u_i$ where $i = 1, 2, \ldots, n$.

Now with the help of the given observations on X and Y and applying the OLS method we can estimate α and β and then we can get $e_i = Y_i - \hat{\alpha} - \hat{\beta} X_i$.

In this way we can obtain the OLS residual terms for each observation.

4.6.1 Spearman's Rank Correlation Test

This test is applicable for large or small samples.

In this method first we have to calculate the rank correlation coefficient between $|e_i|$ and X_i.

This rank correlation coefficient is measured by the formula :

$$r'_{ex} = 1 - \frac{6\sum_{i=1}^{n} D_i}{n(n^2 - 1)}$$

where D_i = difference between the ranks of corresponding pairs of X and $|e|$ and n = number of observations in the sample.

Now we have to test the null hypothesis that the rank correlation coefficient (ρ) is zero against the alternative hypothesis that it is not equal to zero i.e., we have to test $H_0 : \rho = 0$ against the alternative hypothesis, $H_1 : \rho \neq 0$.

The appropriate test statistic is then given by, $t = \frac{r'\sqrt{n-2}}{\sqrt{1 - r'^2}} \sim t_{n-2}$.

This follows a t distribution with $(n - 2)$ degrees of freedom.

Now on the basis of the given sample the null hypothesis $H_0 : \rho = 0$ will be accepted at 5% level of significance if $-t_{0.025}, (n - 2) \leq t \leq t_{0.025}, (n - 2)$ and will be rejected otherwise. If the null hypothesis is accepted then there is no problem of heteroscedasticity. But if it is rejected, then there will be the problem of heteroscedasticity.

4.6.2 Goldfeld and Quandt Test

This test is applicable to large samples. For this test we have to consider the following steps :

(a) We have to order the observations according to the magnitude of the independent variable X.

(b) We have to select an arbitrary number c and omit a certain number of central observations (say c) from the analysis. The remaining observations *i.e.*, $(n - c)$ observations are then divided into two equal parts. Each part, therefore now consists of $\left(\frac{n-c}{2}\right)$ observations. One part includes the small values of X while the other part includes the large values of X.

(c) We fit separate regression lines by OLS procedure to each part and obtain the sum of squared residuals from each of them.

(d) Let $\sum e_1^2$ denote the sum of squared residuals from the sample of low values of X and $\sum e_2^2$ denote the same from the sample of large values of X. Then we calculate

$F = \frac{\sum e_2^2}{\sum e_1^2}$, which will follow the F distribution with $d.f = \left\{\frac{n-c}{2} - k\right\} = \frac{n-c-2k}{2}$ for both numerator and denominator of the ratio. This means that the degrees of freedom of F are $v_1 = \frac{n-c-2k}{2}$, $v_2 = \frac{n-c-2k}{2}$. Here n = total number of observations,

c = number of central observations omitted and k = number of parameters to be estimated including constant term.

Now we have to test the null hypothesis :

H_0 : u_i's are homoscedastic (*i.e.*, the two variances are the same or $H_0 : \sigma_1 = \sigma_2$) against the alternative hypothesis,

H_1 : u_i's are heteroscedastic (*i.e.*, with increasing variances or, $H_1 : \sigma_1 < \sigma_2$).

If the null hypothesis is accepted for the given sample (at a chosen level of significance) then there will be no problem of heteroscedasticity. The null hypothesis will be accepted if $F = \frac{\sum e_2^2}{\sum e_1^2} < F$ (table) and will be rejected if $F = \frac{\sum e_2^2}{\sum e_1^2} > F$ (table). Thus the problem of heteroscedasticity arises if the alternative hypothesis is accepted *i.e.*, null hypothesis is rejected.

4.6.3 Glejser's Test

This test may be outlined as follows :

(a) We perform the regression of Y on all the explanatory variables and we compute the residuals, e's.

(b) We regress the absolute values of e's (*i.e.*, $|e_i|$) on the explanatory variable with which $\sigma_{u_i}^2$ is thought, on *apriori* ground, to be associated.

Let us write the regression equation $|e_i| = \theta_0 + \theta_1 X_i$

Let $\hat{\theta}_0$ and $\hat{\theta}_1$ be the OLS estimators of θ_0 and θ_1.

Now we have to test the null hypothesis $H_0 : \theta_0 = 0$, against the alternative hypothesis, $H_1 : \theta_0 \neq 0$ and $H_0 : \theta_1 = 0$, against the alternative $H_1 : \theta_1 \neq 0$. When H_0 is accepted, there will be no problem of heteroscedasticity. If H_0 is rejected, then there will be the problem of heteroscedasticity. If $\theta_0 = 0$ and $\theta_1 \neq 0$, then there will be exact heteroscedasticity and if $\theta_0 \neq 0$ and $\theta_1 \neq 0$, the case will be referred to as mixed heteroscedasticity.

Example 4.1. A researcher, using time series data for the period 1954-65, estimated the following consumption function : $\hat{C} = -2.945 + 0.9277X$. The following table includes the data used and the residual errors :

Year :	1954	1955	1956	1957	1958	1959	1960	1961	1962	1963	1964	1965
Consumption (C) (billions of \$)	236	254	267	281	290	311	325	335	355	375	401	431
Income (X) (billions of \$)	257	275	293	309	319	337	350	364	385	405	437	469
$e_i = C_i - \hat{C}_i$	0.52	1.82	−1.87	−2.71	−2.99	1.30	3.25	0.26	0.78	2.23	−1.45	−1.14

Note : Here intercept of the consumption function is negative but slope (MPC) is positive and less than unity. Negative intercept does not contradict with the theory as it is linear model and other determinants of consumption have been excluded.

(i) Test for heteroscedasticity, using Spearman's rank correlation coefficient.

(ii) Outline the corrective solution which you would adopt if heteroscedasticity is found significant.

Solution : (i) To apply Spearman's rank correlation test we rank X's and $|e'_i|$ in ascending order. The rankings are shown in the following table. (Table 4.1)

Table 4.1

Value of X (billions of \$)	Rank of X	Value of $\lvert e \rvert$	Rank of $\lvert e \rvert$	D_i	D_i^2
469	1	1.14	9	− 8	64
437	2	1.45	7	− 5	25
405	3	2.23	4	− 1	1
385	4	0.78	10	− 6	36
364	5	0.26	12	− 7	49
350	6	3.25	1	5	25
357	7	1.30	8	− 1	1
319	8	2.99	2	6	36
309	9	2.71	3	6	36
293	10	1.87	5	5	25
275	11	1.82	6	5	25
257	12	0.52	11	1	1
	Total				$\sum D_i^2 = 324$

The rank correlation coefficient estimated from the above data is

$$r'_{xe} = 1 - \frac{6\sum D_i^2}{n(n^2-1)} = 1 - \frac{6 \times 324}{12(12^2-1)}$$

$$= 1 - \frac{1944}{1716} = \frac{1716-1944}{1716} = -\frac{228}{1716} = -0.139.$$

Now we have to test the null hypothesis that the rank correlation coefficient is zero against the alternative hypothesis that it is not equal to zero. *i.e.*, we have to test $H_0 : \rho = 0$ against the alternative $H_1 : \rho \neq 0$. The appropriate test statistic is then given by $t = \frac{r'\sqrt{n-2}}{\sqrt{1-r'^2}} \sim t_{n-2}$ *i.e.*, it follows a 't' distribution with $(n-2)$ degrees of freedom.

Here we have,

$$t = \frac{-0.139 \times \sqrt{12-2}}{\sqrt{1-(0.139)^2}} = \frac{-0.139 \times 3.1623}{\sqrt{0.9806}} = \frac{-0.4396}{0.9902} = -0.444$$

$$\therefore t = -0.444$$

Now the null hypothesis will be accepted for the given sample at 5% level of significance if

$-t_{0.025}, (n-2) \le t \le t_{0.025}, (n-2)$ and will be rejected otherwise. From the table value we see that $t_{0.025}, (n-2) = t_{0.025}, 10 = 2.228$ (as $n = 12$). Here we see that $t = -0.444$ lies in the range -2.228 and 2.228 and hence the null hypothesis will be accepted. We may thus conclude that there is no problem of heteroscedasticity.

(ii) Here we see that heteroscedasticity is not significant. If heteroscedasticity exists then the corrective solution may be outlined as stated below. Let us assume that the pattern of heteroscedasticity is of the form $E(u_t^2) = \sigma_u^2 . x_t$ so that the appropriate transformation of the original model $C_t = \beta\sqrt{x_t} + u_t$ will become $\frac{C_t}{\sqrt{x_t}} = \beta\sqrt{x_t} + \frac{u_t}{\sqrt{x_t}}$.

Here $c_t = C_t - \bar{C}_t$, $x_t = X_t - \bar{X}_t$ and $\bar{u}_t = 0$.

Now by applying OLS to the new variables we can obtain,

$$\hat{\beta} = \frac{\sum c_t}{\sum x_t} = \frac{\bar{c}}{\bar{x}} \text{ and } \hat{\alpha} = \bar{C}_t - \hat{\beta}\bar{X}_t$$

and correspondingly we can estimate $SE(\hat{\alpha})$, $SE(\hat{\beta})$ and the value of R^2.

Example 4.2. The estimated saving function for a 31 year period is given by,

$$\hat{S}_t = -644.1 + 085X_t$$

SE : (117.6) (0.005) $R^2 = 0.903$

After arranging the X's in ascending order and omitting nine central observations we are left with two subsets of data, one with the lower values of X and the other with the higher values of X.

Applying OLS to each subset, we obtained,

(i) For subset 1, $\hat{S}_1 = -738.84 + 0.088\,X$; $R^2 = 0.787$

SE : (189.4) (0.015)

and $\sum e_1^2 = 144,771.5$

(ii) For subset 2, $\hat{S}_2 = 1141.07 + 0.029X$; $R^2 = 0.152$

SE : (709.8) (0.022)

and $\sum e_2^2 = 769,899.2$

By using Goldfeld and Quandt test, examine whether the problem of heteroscedasticity exists or not in this problem.

Solution : For Goldfeld and Quandt test we use the test statistic $F^* = \frac{\sum e_2^2}{\sum e_1^2}$ with $df = v_1 = (n - c - 2k)/2$, $v_2 = (n - c - 2k)/2$

where n = total number of observations,

c = number of central observations omitted and k = no of parameters to be estimated.

Here $n = 31$, $c = 9$, $k = 2$ $\therefore$ $\frac{n-c-2k}{2} = \frac{31-9-4}{2} = \frac{18}{2} = 9$.

Now, $F^* = \frac{\sum e_2^2}{\sum e_1^2} = \frac{769,899.2}{144,771.5} \approx 5$ with $df = (9, 9)$.

From table value we have $F_{0.05}$; 9, 9 = 3.18. Thus we see that $F^* = 5 > F_{0.05}$; 9, 9 = 3.18) and hence the null hypothesis (there is no problem of heteroscedasticity) is rejected at 5% level of significance. Thus the problem is involved with heteroscedasticity.

4.7. Autocorrelation

Let us consider a two-variable linear regression model :

$Y_i = \alpha + \beta X_i + u_i$ for $i = 1, 2, \ldots n$. The model is subject to the problem of autocorrelation if assumption (iv) of CLRM [*i.e.*, $E(u_i, u_j) = cov(u_i, u_j) = 0$] is not satisfied. In other words, the model is subject to the problem of autocorrelation when the disturbance terms are not independent.

i.e., $Cov(u_i, u_j) = E(u_i, u_j) \ne 0$ for $i \ne j$.

Generally, the problem of autocorrelation arises in the case of time series data. Let us consider a model :

$Y_t = \alpha + \beta X_t + u_t$ based on time series data for $t = 1, 2, 3, \ldots r$.

We now consider $u_t = \rho u_{t-1} + \epsilon_t$ with $|\rho| < 1$. This is called first order autoregressive scheme.

where ρ = the coefficient of autocorrelation.

ϵ_t = a random term which fulfils all the usual assumptions of a random variable, that is $E(\epsilon) = 0$, $E(\epsilon^2) = \sigma_\epsilon^2$

and $E(\epsilon_i, \epsilon_j) = 0$ for $i \ne j$.

The complete form of the first order Markov process (the pattern of autocorrelation for all the values of u), is.

$u_t = f(u_{t-1}) = \rho u_{t-1} + \epsilon_t$

$u_{t-1} = f(u_{t-2}) = \rho u_{t-2} + \epsilon_{t-1}$

$u_{t-2} = f(u_{t-3}) = \rho u_{t-3} + \epsilon_{t-2}$

………………………………………

………………………………………

………………………………………

$u_{t-r} = f(u_{t-(r+1)}) = \rho u_{t-(r+1)} + \epsilon_{t-r}$

In order to define the error term in any particular period t we work as follows. We start from the autocorrelation relationship in period t,

$u_t = \rho u_{t-1} + \epsilon_t$ and we perform continuous substitutions of the lagged values of u, as follows.

If we substitute u_{t-1} then we get,

$u_t = \rho[\rho u_{t-2} + \epsilon_{t-1}] + \epsilon_t = \rho^2 u_{t-2} + (\rho\,\epsilon_{t-1} + \epsilon_t)$

If we substitute u_{t-2} then we get,

$$u_t = \rho^2[\rho u_{t-3} + \epsilon_{t-2}] + (\rho\,\epsilon_{t-1} + \epsilon_t)$$

$$= \rho^3 u_{t-3} + (\rho^2\,\epsilon_{t-2} + \rho\,\epsilon_{t-1} + \epsilon_t)$$

If we substitute u_{t-3} then we get,

$u_t = \rho^3[\rho u_{t-4} + \epsilon_{t-3}] + (\rho^2\,\epsilon_{t-2} + \rho\,\epsilon_{t-1} + \epsilon_t)$

$= \rho^4 u_{t-4} + (\rho^3\,\epsilon_{t-3} + \rho^2\,\epsilon_{t-2} + \rho\,\epsilon_{t-1} + \epsilon_t)$

If we continue the substitution process for r periods (where r is very large) we find $u_t = \varepsilon_t + \rho\varepsilon_{t-1} + \rho^2\varepsilon_{t-2} + \rho^3\varepsilon_{t-3} + \ldots\ldots$ (given that as the power of ρ increases to infinity the term with the lagged u, $\rho^r u_{t-r}$, tends to zero, since $|\rho| < 1$).

Thus, $u_t = \sum_{r=0}^{\infty} \rho^r \cdot \varepsilon_{t-r}$

This is the value of the error term when it is autocorrelated with a first order autoregressive scheme.

4.8. Mean, Variance and Covariance of the Autocorrelated Disturbance Variable

1. Mean of the autocorrelated u's

$$E(u_t) = E\sum_{r=0}^{\infty}\rho^r \varepsilon_{t-r} = \sum_{r=0}^{\infty}\rho^r E(\varepsilon_{t-r})$$

But by the assumptions of the distribution of ε we have $E(\varepsilon_{t-r}) = 0$. Therefore $E(u_t) = 0$.

2. Variance of the autocorrelated u's

By definition var $(u_t) = E[u_t - E(u_t)]^2$

$= E\left[u_t^2\right]$ as $E(u_t) = 0$

$$= E\left[\sum_{r=0}^{\infty}\rho^r \varepsilon_{t-r}\right]^2$$

$= \sum(\rho^r)^2 E(\varepsilon_{t-r})^2 = \sum(\rho^r)^2 \text{var}(\varepsilon_{t-r}) = \sum\rho^{2r}\sigma_\varepsilon^2$

$= \sigma_\varepsilon^2(1 + \rho^2 + \rho^4 + \rho^6 + \ldots)$

$\therefore E(u_t^2) = \sigma_\varepsilon^2(1 + \rho^2 + \rho^4 + \rho^6 + \ldots)$ [Follows from the sum of G.P. series]

or, $\text{var}(u_t) = \dfrac{\sigma_\varepsilon^2}{1-\rho^2}$ for $|\rho| < 1$.

3. Covariance of the autocorrelated u's

Since $u_t = \varepsilon_t + \rho\varepsilon_{t-1} + \rho^2\varepsilon_{t-2} + \ldots$

and $u_{t-1} = \varepsilon_{t-1} + \rho\varepsilon_{t-2} + \rho^2\varepsilon_{t-3} + \ldots\ldots$

Now, $\text{cov}(u_t, u_{t-1}) = E\{[u_t - E(u_t)][u_{t-1} - E(u_{t-1})]\}$

$= E[u_t u_{t-1}]$ as $\begin{cases} E(u_t) = 0 \\ E(u_{t-1}) = 0 \end{cases}$

$= E\left[\left(\varepsilon_t + \rho\varepsilon_{t-1} + \rho^2\varepsilon_{t-2} + \ldots\right)\left(\varepsilon_{t-1} + \rho\varepsilon_{t-2} + \rho^2\varepsilon_{t-3} + \ldots\right)\right]$

$= E\left[\{\varepsilon_t + \rho(\varepsilon_{t-1} + \rho\varepsilon_{t-2} + \ldots)\}\left(\varepsilon_{t-1} + \rho\varepsilon_{t-2} + \rho^2\varepsilon_{t-3} + \ldots\right)\right]$

$= E\left[(\varepsilon_t)\left(\varepsilon_{t-1} + \rho\varepsilon_{t-2} + \rho^2\varepsilon_{t-3} + \ldots\right)\right] + E\left[\rho(\varepsilon_{t-1} + \rho\varepsilon_{t-2} + \ldots)^2\right]$

$= 0 + \rho E(\varepsilon_{t-1} + \rho\varepsilon_{t-2} + \ldots)^2$

$= \rho E(\varepsilon_{t-1}^2 + \rho^2\varepsilon_{t-2}^2 + \ldots + \text{cross products})$

$= \rho\left(\sigma_\varepsilon^2 + \rho^2\sigma_\varepsilon^2 + \ldots + 0\right)$

$= \rho\left[\sigma_\varepsilon^2\left(1 + \rho^2 + \rho^4 + \rho^6 + \ldots\right)\right]$

$= \rho\sigma_\varepsilon^2 \dfrac{1}{1-\rho^2}$ for $|\rho| < 1$

$\therefore \text{cov}(u_t, u_{t-1}) = \rho\sigma_u^2 \qquad \left[\because \text{var}(u_t) = \dfrac{\sigma_\varepsilon^2}{1-\rho^2} = \sigma_u^2\right]$

Similarly, $\text{cov}(u_t, u_{t-2}) = E(u_t u_{t-2}) = \rho^2\sigma_u^2$ and in general $\text{cov}(u_t, u_{t-s}) = \rho^s\sigma_u^2$ (for $s \neq t$). Summarising on the basis of (1), (2) and (3), we find that when u_t's are autocorrelated, then $u_t \sim N\left(0, \dfrac{\sigma_\varepsilon^2}{1-\rho^2}\right)$ and $\text{cov}(u_t, u_{t-1}) \neq 0$.

4.9. Consequences of Autocorrelation

When the disturbance term exhibits serial correlation the value as well as the standard errors of the parameter estimates are affected. In particular :

1. OLS estimates are unbiased

In the deviation form of the simple regression model we get $\hat\beta = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$ where $x_i = X_i - \bar X$ and $y_i = Y_i - \bar Y$.

or, $\hat\beta = \dfrac{\sum_{i=1}^{n} x_i(\beta x_i + u_i)}{\sum_{i=1}^{n} x_i^2}$ or, $\hat\beta = \beta + \dfrac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2}$ $\qquad \left[\begin{array}{l} \because Y_i = \alpha + \beta X_i + u_i \\ \text{or, } \bar Y = \alpha + \beta\bar X, \text{ as } \bar u = 0 \\ \text{or, } Y_i - \bar Y = \beta(X_i - \bar X) + u_i \\ \text{or, } y_i = \beta x_i + u_i \end{array}\right.$

Now, $E(\hat\beta) = \beta + \dfrac{\sum_{i=1}^{n} x_i E(u_i)}{\sum_{i=1}^{n} x_i^2}$

$$\therefore \text{ Bias in } \hat{\beta} = E(\hat{\beta}) - \beta = \frac{\sum_{i=1}^{n} x_i E(u_i)}{\sum_{i=1}^{n} x_i^2} = 0 \quad [\because E(u_i) = 0]$$

Hence bias in $\hat{\beta} = 0$.

Thus, irrespective of whether the residuals are serially independent or not, the estimates of the parameters have no statistical bias, so long as the u's and X's are uncorrelated.

2. The variances of OLS estimates are underestimated

The variance of estimate $\hat{\beta}$ in simple regression model will be biased downwards (i.e., underestimated) when u's are autocorrelated.

We have seen that [2.9 (3) i.e. from the property of unbiasedness]

$$\text{var}(\hat{\beta}) = E\left[\sum k_i u_i\right]^2$$

$$= E\left[\sum k_i^2 u_i^2 + 2\sum k_i k_j u_i u_j\right]$$

$$= \sum k_i^2 E(u_i^2) + 2\sum k_i k_j E(u_i, u_j)$$

Replacing Σk_i and $\Sigma k_i^2 \left[\because k_i = \frac{x_i}{\Sigma x_i^2}\right]$ we get,

$$\text{var}(\hat{\beta}) = \Sigma\left[\frac{x_i}{\Sigma x_i^2}\right]^2 \cdot E(u_i^2) + 2\Sigma\left[\frac{x_i x_j}{\Sigma x_i^2 \cdot \Sigma x_j^2}\right] E(u_i, u_j)$$

If $E(u_i, u_j) = 0$, then, $\text{var}(\hat{\beta}) = \sigma_u^2 \Sigma k_i^2 = \frac{\sigma_u^2}{\Sigma x_i^2}$

$$\left[\because k_i = \frac{x_i}{\Sigma x_i^2} \text{ and } \Sigma k_i^2 = \Sigma\left(\frac{x_i}{\Sigma x_i^2}\right)^2 = \frac{\Sigma x_i^2}{(\Sigma x_i^2)^2} = \frac{1}{\Sigma x_i^2}\right]$$

However, with u's related with a first order autoregressive scheme we have

$$E(u_t^2) = \sigma_u^2 = \frac{\sigma_\varepsilon^2}{1-\rho^2} \text{ and } E(u_t\, u_{t-s}) = \rho^s \sigma_u^2.$$

Therefore, $\text{var}(\hat{\beta}) = \sigma_u^2 \frac{1}{\Sigma x_i^2} + 2\sigma_u^2 \Sigma \frac{x_i x_j}{(\Sigma x_i^2)^2} \cdot \rho^s$. Expanding the second term in the above expression, we obtain :

$$\text{var}(\hat{\beta}) = \frac{\sigma_u^2}{\Sigma x_i^2} + 2\sigma_u^2 \left\{\rho \frac{\sum_{i}^{n-1} x_i x_{i+1}}{\left(\sum_{i=1}^{n} x_i^2\right)^2} + \rho^2 \frac{\sum_{i}^{n-2} x_i x_{i+2}}{\left(\sum_{i=1}^{n} x_i^2\right)^2} + \cdots\right\}$$

$$\text{or} \quad \text{var}(\hat{\beta}) = \frac{\sigma_u^2}{\Sigma x_i^2}\left\{1 + 2\rho \frac{\sum_{i}^{n-1} x_i x_{i+1}}{\left(\sum_{i=1}^{n} x_i^2\right)^2} + 2\rho^2 \frac{\sum_{i}^{n-2} x_i x_{i+2}}{\left(\sum_{i=1}^{n} x_i^2\right)^2} + \cdots + 2\rho^{n-1} \frac{x_1 x_n}{\sum_{i}^{n} x_i^2}\right\}$$

In the absence of autocorrelation, $\text{var}(\hat{\beta}) = \frac{\sigma_u^2}{\Sigma x_i^2}$. But presently, in case ρ is positive (case of positive autocorrelation) and if X is also positively correlated ($\Sigma x_i x_j \neq 0$), the expression in the bracket is almost certainly greater than unity. This, in turn, proves that estimate of variance, in most cases, will have downward bias due to positive autocorrelation. In case the explanatory variable X of the model is random, the covariance of successive values is zero ($\Sigma x_i x_j = 0$). Under such circumstances the bias in $\text{var}(\hat{\beta})$ will not be serious even though u is autocorrelated.

3. The Predictions will be inefficient

If the values of u are autocorrelated, the predictions based on least squares estimates will be inefficient, in the sense that they will have a larger variance as compared with predictions based on estimates obtained from other econometric techniques (like GLS).

4.10. Test for Autocorrelation

There are various ways of testing autocorrelation. The traditionally applied tests are the Durbin-Watson test and the Von Neumann ratio method.

4.10.1 Durbin-Watson Test

J. Durbin and G.S. Watson have suggested a test which is applicable to small samples. However, the test is appropriate only for the first order autoregressive scheme :

$$u_t = \rho u_{t-1} + \varepsilon_t$$

In this method we are interested in testing the null hypothesis $H_0 : \rho = 0$ (or H_0 : the u's are not autocorrelated with a first order scheme) against the alternative $H_1 : \rho \neq 0$ (or H_1 : the u's are autocorrelated with a first order scheme).

To test the null hypothesis we use the Durbin-Watson statistic, given by,

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \text{ where } e_t \text{ for } t = 1, 2, n \text{ are the OLS residual terms.}$$

$$\text{Now, } d = \sum_{t=2}^{n}(e_t - e_{t-1})^2 \Big/ \sum_{t=1}^{n} e_t^2 = \frac{\sum_{t=2}^{n} e_t^2 + \sum_{t=2}^{n} e_{t-1}^2 - 2\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2}$$

For very large values of n (for large samples) $\sum_{t=1}^{n} e_t^2, \sum_{t=2}^{n} e_{t-1}^2$ and $\sum_{t=2}^{n} e_t^2$ are approximately equal.

$$\text{Then, } d \approx \frac{2\sum_{t=2}^{n} e_{t-1}^2 - 2\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e_{t-1}^2} \quad \therefore d \approx 2\left(1 - \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e_{t-1}^2}\right)$$

But $\hat{\rho} = \dfrac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e_{t-1}^2}$ where $\hat{\rho}$ is the OLS estimator of ρ in the model $e_t = \rho e_{t-1} + \varepsilon_t$ for $t = 2, 3, \ldots n$ $\therefore d \approx 2(1 - \hat{\rho})$

Since $-1 < \hat{\rho} < 1$.

$\therefore 0 < d < 4$.

This shows that d lies between 0 and 4.

Firstly, if there is no autocorrelation $\hat{\rho} = 0$ and $d = 2$. Thus if from the sample data we find d (observed) $= d^* = 2$, we accept that there is no autocorrelation in the function.

Secondly, if $\hat{\rho} = +1$, $d = 0$ and we have perfect-positive autocorrelation.

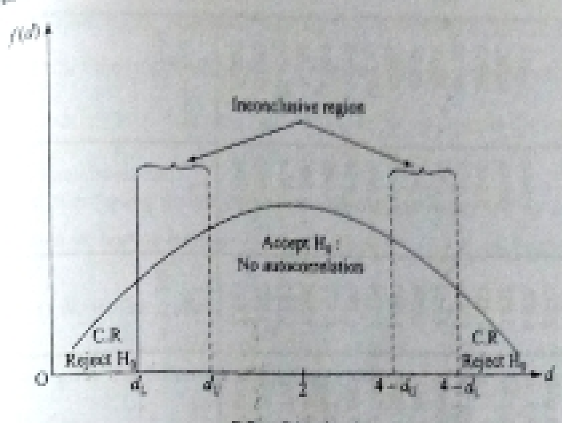
Thirdly, if $\hat{\rho} = -1$, $d = 4$ we have perfect negative autocorrelation. Therefore if $2 < d^* < 4$ there is some degree of negative autocorrelation, which is stronger if the value of d^* is higher.

If there is no problem of autocorrelation $\hat{\rho}$ should be zero and $d = 2$.

Thus we have to test the null hypothesis $H_0 : d = 2$ against the alternative $H_1 : d \neq 2$.

Here the problem is that the exact sampling distribution of the statistic 'd' is not known. What Durbin and Watson have done is to specify one upper limit and one lower limit of d.

Let d_U stand for the upper limit of d and d_L stand for the lower limit of d. This is shown in the diagram on the next page. (Fig. 4.1)



C.R ⇒ Critical region

Fig. 4.1. The critical regions of d test are shown

Now with the help of d_L and d_U we have to determine whether autocorrelation exists or not because the values of d_L and d_U are available at the 5% and 1% levels of significance.

We calculate $d^* = 2(1 - \hat{\rho})$, $\hat{\rho} = \dfrac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e_{t-1}^2}$ from the sample.

We may now consider the following cases :

1. If it is found that $d^* < d_L$ we reject the null hypothesis of no autocorrelation and accept that there is positive autocorrelation of first order.
2. If $d^* > (4 - d_L)$ we reject the null hypothesis of no autocorrelation and accept that there is negative autocorrelation of the first order.
3. If $d_U < d^* < (4 - d_L)$ we accept the null hypothesis of no autocorrelation.
4. If $d_L < d^* < d_U$ or if $(4 - d_U) < d^* < (4 - d_L)$ the test is inconclusive.

Limitations of Durbin-Watson Test

There are some limitations of Durbin-Watson test.

(i) There exist inconclusive regions and hence if the value of d^* lies between either d_L and d_U or between $4 - d_U$ and $4 - d_L$, then we cannot conclude whether autocorrelation exists or not.

(ii) This test method is appropriate only when the nature of the autocorrelation is of first order autoregressive type. But it is not appropriate when autocorrelation is of higher order and non linear type.

(iii) If there is any lagged variable as independent variable in the model, then Durbin-Watson statistic 'd' is inappropriate in testing for autocorrelation.

Table 4.2

Y_t	X_t	$y_t = Y_t - \bar{Y}$	y_t^2	$x_t = X_t - \bar{X}$	x_t^2	$x_t y_t$	$\hat{Y}_t = \hat{\alpha} + \hat{\beta} X_t$	$e_t = Y_t - \hat{Y}_t$	e_t^2	$e_t - e_{t-1}$	$(e_t - e_{t-1})^2$
2	1	– 5	25	– 7	49	35	0.322	1.678	2.815	–	–
1	2	– 6	36	– 6	36	36	1.276	– 0.276	0.076	– 1.954	3.818
2	3	– 5	25	– 5	25	25	2.230	– 0.230	0.052	0.046	0.002
3	4	– 4	16	– 4	16	16	3.184	– 0.184	0.033	0.046	0.002
3	5	– 4	16	– 3	9	12	4.138	– 1.138	1.295	– 0.954	0.910
2	6	– 5	25	– 2	4	10	5.092	– 3.092	9.560	– 1.954	3.818
5	7	– 2	4	– 1	1	2	6.046	– 1.046	1.094	2.046	4.186
6	8	– 1	1	0	0	0	7.000	– 1.000	1.000	0.046	0.002
11	9	4	16	1	1	4	7.954	3.046	9.278	4.046	16.370
10	10	3	9	2	4	6	8.908	1.092	1.192	– 1.954	3.818
12	11	5	25	3	9	15	9.862	2.138	4.571	1.046	1.094
15	12	8	64	4	16	32	10.816	4.184	17.505	2.046	4.186
10	13	3	9	5	25	15	11.770	– 1.770	3.132	– 5.594	35.450
11	14	4	16	6	36	24	12.724	– 1.724	2.972	0.046	0.002
12	15	5	25	7	49	35	13.678	– 1.678	2.815	0.046	0.002
$\Sigma Y_t = 105$ $\therefore \bar{Y} = \frac{\Sigma Y_t}{n} = \frac{105}{15} = 7$	$\Sigma X_t = 120$ $\therefore \bar{X} = \frac{\Sigma X_t}{n} = \frac{120}{15} = 8$	$\Sigma y_t = 0$	$\Sigma y_t^2 = 312$	$\Sigma x_t = 0$	$\Sigma x_t^2 = 280$	$\Sigma x_t y_t = 267$	–	–	$\Sigma e_t^2 = 57.390$	–	$\Sigma (e_t - e_{t-1})^2 = 73.660$

Example 4.3. Consider the model ; $Y_t = \alpha + \beta X_t + u_t$ with the following observations of Y and X :

X	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Y	2	1	2	3	3	2	5	6	11	10	12	15	10	11	12

Test for autocorrelation.

Solution : For testing the autocorrelation first we have to estimate the regression model and then we have to estimate $e_t = Y_t - \hat{Y}_t$ where $\hat{Y}_t = \hat{\alpha} + \hat{\beta} X_t$ and finally we have to find the D.W. statistic $d^* = \frac{\Sigma (e_t - e_{t-1})^2}{\Sigma e_t^2}$

Table 4.2 gives the necessary calculations to estimate $\hat{\alpha}$ and $\hat{\beta}$ in the model : $Y_t = \alpha + \beta X_t + u_t$

Now we have, $\hat{\beta} = \frac{\Sigma x_t y_t}{\Sigma x_t^2} = \frac{267}{280} = 0.954$

$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 7 - 0.954 \times 8 = 7 - 7.632 = -0.632$

$\hat{\sigma}_u^2 = \frac{\Sigma e_t^2}{n-2} = \frac{57.390}{13} = 4.4146$

$\text{var}(\hat{\beta}) = \frac{\hat{\sigma}_u^2}{\Sigma x_t^2} = \frac{4.4146}{280} = 0.0157$

$\therefore SE(\hat{\beta}) = \sqrt{0.0157} = 0.1252$

$\text{var}(\hat{\alpha}) = \hat{\sigma}_u^2 \frac{\Sigma X_t^2}{n \Sigma x_t^2} = \frac{4.4146 \times 1240}{15 \times 280} = \frac{5474.104}{4200} = 1.3036.$

where $\Sigma X_t^2 = 1^2 + 2^2 + ... + 15^2 = 1240$

$\therefore SE(\hat{\alpha}) = \sqrt{1.3036} = 1.1417.$

$R^2 = \frac{\hat{\beta}^2 \Sigma x_t^2}{\Sigma y_t^2} = \frac{(0.954)^2 \times 280}{312} = \frac{254.8324}{312} = 0.816$

Thus the estimated model is :

$\hat{Y}_t = -0.632 + 0.954\, X_t \qquad R^2 = 0.816$

SE : (1.1417) (0.1252)

Now Durbin-Watson statistic $d^* = \frac{\Sigma (e_t - e_{t-1})^2}{\Sigma e_t^2} = \frac{73.660}{57.390} = 1.283$

Values of d_L and d_U at 5% level of significance, with $n = 15$ and one explanatory variable ($K' = 1$), are $d_L = 1.08$ and $d_U = 1.36$. Here we see that $d_L < d^* < d_U$ and hence the test is inconclusive.

In other words on the basis of Durbin–Watson test we cannot say whether autocorrelation problem exists or not.

4.10.2. Von Neumann Ratio Method of Testing Autocorrelation

This is the ratio of the variance of the first differences of any variable X over the variance of X and is given by,

$$VN = \frac{\sum_{t=2}^{n}(X_t - X_{t-1})^2/(n-1)}{\Sigma(X_t - \bar{X})^2/n}$$

The Von Neumann ratio is applicable for directly observed series and for variables which are random, that is, variables whose successive values are not autocorrelated. In the case of the u's, their values are not directly observable but are estimated from the OLS residuals (e's). For large samples ($n > 30$) we can write the Von Neumann ratio as,

$$VN = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2/(n-1)}{\Sigma(e_t - \bar{e})^2/n}$$

This test statistic is used for ensuring the existence of autocorrelation.

This procedure is, however, not applicable for testing the autocorrelation of the u's, especially if the sample is small ($n < 30$).

4.11. Methods for Estimating Regression Parameters in the Presence of the Problem of Autocorrelation

Once the autocorrelation is detected, the appropriate corrective procedure is : (i) to obtain an estimate of autocorrelation coefficient ρ, (ii) the original data is transformed and (iii) OLS is applied to the transformed data.

Let our model be given by, $Y_t = \alpha + \beta X_t + u_t$ .... (1)

where $u_t = \rho u_{t-1} + \varepsilon_t$ with $|\rho| < 1$.

If we take a lagged form of the model (1) and multiply both sides by ρ we obtain,

$\rho Y_{t-1} = \rho\alpha + \rho\beta X_{t-1} + \rho u_{t-1}$ .... (2)

Now subtracting (2) from (1) we get,

$$Y_t - \rho Y_{t-1} = \alpha(1-\rho) + \beta(X_t - \rho X_{t-1}) + (u_t - \rho u_{t-1})$$

or, $Y_t^* = \alpha^* + \beta(X_t - \rho X_{t-1}) + \varepsilon_t$ $[\because u_t = \rho u_{t-1} + \varepsilon_t \quad \varepsilon_t = u_t - \rho u_{t-1}]$

or, $Y_t^* = \alpha^* + \beta X_t^* + \varepsilon_t$ .... (3)

[where $Y_t^* = Y_t - \rho Y_{t-1}$, $X_t^* = X_t - \rho X_{t-1}$ and $\alpha^* = \alpha(1-\rho)$]

Here ε_t satisfies all the properties of CLRM.

It should be noted that in transforming (1) into (3), one observation shall be lost because of lagging and subtracting (2) from (1) we can apply OLS to the transformed relation (3) to obtain $\hat{\alpha}^*$ and $\hat{\beta}$ for our parameters α and β in (1).

$\hat{\beta} = \sum_{t=2}^{n}(X_t^* - \bar{X}_t^*)(Y_t^* - \bar{Y}_t^*) \Big/ \sum_{t=2}^{n}(X_t^* - \bar{X}_t^*)^2$ and $\hat{\alpha} = \frac{\hat{\alpha}^*}{1-\rho}$. It can be shown that

$\text{var}(\hat{\alpha}) = \left(\frac{1}{1-\rho}\right)^2 \cdot \text{var}(\hat{\alpha}^*)$ because $\hat{\alpha}$ is perfectly and linearly related to $\hat{\alpha}^*$. Again since ε_t satisfies all standard assumptions of CLRM, the variances of $\hat{\alpha}^*$ and $\hat{\beta}$ would be given by our standard (OLS) formulae,

$$\text{var}(\hat{\alpha}^*) = \frac{\sigma_u^2 \sum_{t=2}^{n} X_t^{*2}}{n^* \sum_{t=2}^{n}(X_t^* - \bar{X}_t^*)^2}$$

$\text{var}(\hat{\beta}) = \frac{\sigma_u^2}{\sum_{t=2}^{n}(X_t^* - \bar{X}_t^*)^2}$, where $n^* = n - 1$, since one observation is lost in lagging and subtracting to obtain X_t^*.

Estimators obtained from (3) are efficient, only if our sample size is large so that loss of one observation becomes negligible.

The above procedure is applicable only if the value of ρ is known.

Now we have to examine the methods through which ρ can be estimated.

Method I : Using extraneous information on ρ

Sometimes an investigator can make some reasonable guess about the value of the autoregressive coefficient by using his knowledge or intuition about the relationship under study. Usually we assume $\rho = 1$. Now the transformed model becomes $(Y_t - Y_{t-1}) = \beta(X_t - X_{t-1}) + \varepsilon_t$

where $\varepsilon_t = u_t - u_{t-1}$

In this method we can estimate only β but α cannot be estimated.

But if $\rho = -1$, then α and β can be estimated at a time.

Method II : Estimation of ρ from the d-statistic

From the Durbin-Watson test statistic we know that $d \approx 2(1-\hat{\rho})$

where $\hat{\rho} = \sum_{t=2}^{n} e_t e_{t-1} \Big/ \sum_{t=2}^{n} e_{t-1}^2$

Suppose that we calculate certain value of d statistic $= d^*$ from sample data,

then $d^* = 2(1-\hat{\rho})$ so that $\hat{\rho} = 1 - \frac{1}{2}d^*$.

When ρ is estimated then we can estimate $\hat{\alpha}$ and $\hat{\beta}$ from the model.

It should be noted that $\hat{\rho}$ will not be accurate if the sample size is small. This is true only for large samples.

Method III : Iterative procedure

In this method we remove autocorrelation gradually starting from the simplest possible form of a first order scheme. First we obtain the residuals and apply OLS to them :

$e_t = \rho e_{t-1} + v_t$

With the estimated $\hat{\rho}$ from above relation we transform the original data and then apply OLS to the model :

$(Y_t - \hat{\rho} Y_{t-1}) = (1 - \hat{\rho})\alpha + \beta(X_t - \hat{\rho} X_{t-1}) + u_t^*$

where $u_t^* = (u_t - \hat{\rho} u_{t-1})$

Once again we apply Durbin-Watson test to examine whether autocorrelation persists. If it does, then we once again apply OLS to the newly obtained e_t^* from the above transformed data :

$e_t^* = \rho e_{t-1}^* + v_t$

We use new estimate $\hat{\hat{\rho}}$ to transform the original observations and so on. We keep proceeding until a set of non autocorrelated residuals are obtained.

Example 4.4. Simple consumption function is estimated from the hypothetical data on income (Y_{dt}) given in Table 4.3. It is given that the estimated consumption function is : $\hat{C}_t = 3.29 + 0.9067 Y_{dt}$, $R^2 = 0.99$

SE : (0.0055)

(i) Examine whether any autocorrelation problem exists or not. (ii) If the problem of autocorrelation exists in the disturbance terms, then assuming a first order autoregressive scheme of the form : $e_t = \rho e_{t-1} + v_t$ and estimate the transformed regression model (consumption function).

Solution : (i) The required calculations are shown in Table 4.3

The estimated consumption function is given here by :

$\hat{C}_t = 3.29 + 0.9067 Y_{dt}$, $R^2 = 0.99$

SE : (0.0055)

This equation explains almost all variations in consumption (99%) ; but the variance of $\hat{\beta}$ is extremely small.

Now we have to examine the error terms to see whether the evidence of autocorrelation exists or not. In order to do so we find the value of the Durbin-Watson statistic:

$$d^* = \frac{\Sigma(e_t - e_{t-1})^2}{\Sigma e_t^2} = \frac{142.08}{143.14} = 0.9926 \text{ (after substitution of the values from Table 4.3).}$$

At 5% level of significance for $n = 19$, $k' = 1$ (for one explanatory variable), $d_L = 1.18$ and $d_U = 1.40$. Here we see that $d^* < d_L$ (= 1.18) and hence we reject the null hypothesis of no autocorrelation in favour of the alternative hypothesis of (positive) autocorrelated disturbance terms.

(ii) Assuming a first order autoregressive scheme of the form : $e_t = \rho e_{t-1} + v_t$, we may estimate the value of ρ, given

by
$$\hat{\rho} = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e_{t-1}^2} = \frac{67.99}{140.25} = 0.4848$$

With this estimated value of ρ, we obtain transformed consumption function

$C_t^* = \alpha^* + \beta Y_{dt}^* + e_t$

where $C_t^* = C_t - \hat{\rho} C_{t-1} = C_t - 0.48 C_{t-1}$

$Y_t^* = Y_t - \hat{\rho} Y_{dt-1} = Y_{dt} - 0.48 Y_{dt-1}$

$\alpha^* = \alpha(1 - \hat{\rho})$

The relevant computations are shown in Table 4.4.

Now we have to regress C_t^* on Y_{dt}^* to get the estimates of the parameters in

$C_t^* = \alpha^* + \beta Y_{dt}^* + e_t$

From Table 4.4 it can be shown that

$\Sigma c_t^* y_{dt}^* = 40755.45$, $\Sigma y_{dt}^{*2} = 44969.84$ where $c_t^* = C_t - \bar{C}^*$ and $y_{dt}^* = Y_{dt}^* - \bar{Y}^*$

$\bar{C}^* = \frac{\Sigma C_t^*}{n} = 163.74$, $\bar{Y}^* = \frac{\Sigma Y_{dt}^*}{n} = 178.58$

and $\hat{\rho} = 0.48$

Now, $\hat{\beta} = \frac{\Sigma c_t^* y_{dt}^*}{\Sigma y_{dt}^{*2}} = \frac{40755.45}{44969.84} = 0.9062$

$\hat{\alpha}^* = \bar{C}^* - \hat{\beta} \bar{Y}^*$

$= 163.74 - 0.9062 \times 178.58$

$= 163.74 - 161.83 = 1.91$

$\therefore \hat{\alpha} = \frac{\hat{\alpha}^*}{1 - \hat{\rho}} = \frac{1.91}{1 - 0.48} = \frac{1.91}{0.52} = 3.673$

Thus the regression model in the transformed form is :

$\hat{C}_t^* = 1.91 + 0.9062 Y_{dt}^*$

which may now be stated in terms of the original variables as $\hat{C}_t = 3.673 + 0.9062 Y_{dt}$ (where $n = 18$).

4.12. Estimation in Levels versus First Differences

If the D.W. test rejects the hypothesis of zero serial correlation, what is the next step?

In such cases we may estimate a regression by transforming all the variables by ρ differencing, that is, regress $Y_t - \hat{\rho} Y_{t-1}$ on $X_t - \hat{\rho} X_{t-1}$ where $\hat{\rho}$ is the estimated ρ. However, since $\hat{\rho}$ is subject to sampling errors, one other alternative that is followed if the D.W. statistic d is small is to use a first difference equation. In fact a rough rule

Table 4.3

Calculation of d-statistic for $\hat{C}_t = 3.39 + 0.966\, Y_d$

Year	Consumption Expenditure C_t (₹)	Disposable income Y_d (₹)	Estimated consumption $\hat{C}_t = 3.39 + 0.966 Y_d$	e_t	e_t^2	e_{t-1}	e_{t-1}^2	$e_t - e_{t-1}$	$(e_t - e_{t-1})^2$	$e_t e_{t-1}$
1951	206.3	226.6	208.6	−2.3	5.29	–	–	–	–	–
1952	216.7	238.6	219.2	−2.5	6.25	−2.3	5.29	−0.2	0.04	5.75
1953	230.0	252.6	232.1	−2.1	4.41	−2.5	6.25	0.4	0.16	5.25
1954	236.5	257.4	236.5	0.00	0.00	−2.1	4.41	2.1	4.41	0.00
1955	254.4	275.3	252.7	1.7	2.89	0.0	0.00	1.7	2.89	0.00
1956	266.7	293.2	268.9	−2.2	4.84	1.7	2.89	−3.9	15.21	−3.74
1957	281.4	308.5	282.8	−1.4	1.96	−2.2	4.84	0.8	0.64	3.08
1958	290.1	318.8	292.1	−2.0	0.04	−1.4	1.96	−0.6	0.36	2.80
1959	311.2	337.3	308.9	2.3	5.29	−2.0	4.00	4.3	18.49	−4.60
1960	325.2	350.0	320.4	4.8	23.04	2.3	5.29	2.5	6.25	11.04
1961	335.2	364.4	333.4	1.8	3.24	4.8	23.04	−3.0	9.00	8.64
1962	355.1	385.5	352.4	2.7	7.29	1.8	3.24	0.9	0.81	4.86
1963	375.0	404.6	369.9	5.1	26.01	2.7	7.29	2.4	5.76	13.77
1964	401.2	438.1	400.2	1.0	1.00	5.1	26.01	−4.1	16.81	5.1
1965	432.8	473.2	432.0	0.8	0.64	1.0	1.00	−0.2	0.04	0.8
1966	466.3	511.9	467.1	−0.8	0.64	0.8	0.64	−1.6	2.56	−0.64
1967	492.1	456.3	498.2	−6.1	37.21	−0.8	0.64	−5.3	28.09	4.88
1968	536.2	591.0	538.7	−2.5	6.25	−6.1	37.21	3.6	12.96	15.25
1969	579.6	634.2	577.9	1.7	2.89	−2.5	6.25	4.2	17.64	−4.25
Total					$\Sigma e_t^2 = 143.14$		$\Sigma e_{t-1}^2 = 140.25$		$\Sigma(e_t - e_{t-1})^2 = 142.08$	$\Sigma e_t e_{t-1} = 67.98$

Table 4.4

Calculation of transformed variables C_t^* and Y_t^*

Year	C_t	$0.48\, C_t$	$0.48\, C_{t-1}$	$C_t^* = C_t - 0.48\, C_{t-1}$	Y_{dt}	$0.48\, Y_{dt}$	$0.48\, Y_{dt-1}$	$Y^*_{dt} = Y_{dt} - 0.48\, Y_{dt-1}$
1951	206.3	99.02	104.01	102.29	226.6	108.76	114.38	112.22
1952	216.7	104.01	110.40	106.30	238.3	114.38	121.24	117.06
1953	230.0	110.40	113.52	116.48	252.6	121.24	123.55	129.05
1954	236.5	113.52	122.11	114.39	257.4	123.55	132.14	125.26
1955	254.4	122.11	128.01	126.39	275.3	132.14	140.73	134.57
1956	266.7	128.01	135.07	131.63	293.2	140.73	148.08	145.12
1957	281.4	135.07	139.24	142.16	308.5	148.08	153.02	155.48
1958	290.1	139.24	149.37	140.73	318.8	153.02	161.90	156.90
1959	311.2	149.37	156.09	155.11	337.3	161.90	168.00	169.30 Here n = 18
1960	325.2	156.09	160.89	164.31	350.0	168.00	174.90	175.10
1961	335.2	160.89	170.44	164.76	364.4	174.90	185.04	179.36
1962	355.1	170.44	180.00	175.10	385.5	185.04	194.20	191.30
1963	375.0	180.00	192.57	182.43	404.6	194.20	210.28	194.32
1964	401.2	192.57	207.74	193.46	438.1	210.28	227.13	210.97
1965	432.8	207.74	223.82	208.98	473.2	227.13	245.71	226.49
1966	466.3	223.82	236.20	230.10	511.9	245.71	262.22	249.68
1967	492.1	236.20	257.37	234.73	546.3	262.22	283.68	262.62
1968	536.2	257.37	278.20	258.00	591.0	283.68	304.41	286.59
1969	579.6	278.20	–	–	634.2	304.41	–	
Total				$\bar{C}^* = \Sigma C_t^*/n = 163.74$				$\bar{Y}^* = \Sigma Y_{dt}^*/n = 178.98$

is : Estimate an equation on first differences whenever the D.W. statistics is < R^2. In first difference equations, we regress $(Y_t - Y_{t-1})$ on $(X_t - X_{t-1})$ (with all explanatory variables differences similarly). The implicit assumption is that the first differences of the errors $(u_t - u_{t-1})$ are independent. For instance, if $Y_t = \alpha + \beta X_t + u_t$ is the regression equation, then $Y_{t-1} = \alpha + \beta X_{t-1} + u_{t-1}$ and we have by subtraction $(Y_t - Y_{t-1}) = \beta(X_t - X_{t-1}) + (u_t - u_{t-1})$

If the errors in this equation are independent, we can estimate the equation by OLS. However, since the constant term α disappears after subtraction, we should be estimating the regression equation with no constant term. Often, we find a constant term also included in regression equation with first differences. This procedure is valid only if there is a linear trend term in the original equation.

4.13. Multicollinearity–Meaning and Sources

Very often, the data we use in multiple regression analysis cannot give decisive answers to the questions we pose. This is because the standard errors may be very high or the 't' ratios may be very low. Confidence intervals for the parameters of interest may thus be very wide. This sort of situation occurs when explanatory variables display little variation and / or high intercorrelations. The situation where the explanatory variables are highly intercorrelated is referred to as multicollinearity. When the explanatory variables are highly intercorrelated, it becomes difficult to disentangle the separate effects of each of the explanatory variables on the explained variable.

If the explanatory variables are perfectly linearly correlated, that is, if the correlation coefficient of those variables is equal to unity (i.e. $r_{X_1X_2} = 1$) the parameters become indeterminate : it is impossible to obtain numerical values for each parameter separately and the method of least squares breaks down. At the other extreme if the explanatory variables are not intercorrelated at all (i.e., the correlation coefficient for all these variables is equal to zero), the variables are called **orthogonal** and there are no problems concerning the estimates of the coefficients, at least so far as multicollinearity is concerned.

Actually, in the case of orthogonal X's, there is no need to perform a multiple regression analysis : each parameter, β_i, can be estimated by a simple regression of Y on the corresponding explanatory variable X_i : $Y = f(X_i)$.

In practice, neither of the above two extreme cases (of orthogonal X's or perfect collinear X's) is often met.

In most cases there is some degree of intercorrelation among the explanatory variables, due to the interdependence of many economic magnitudes over time.

Multicollinearity is not a condition that either exists or does not exist in economic functions, but rather a phenomenon inherent in most relationships due to the nature of economic magnitudes. There is no conclusive evidence concerning the degree of collinearity which, if present, will affect seriously the parameter estimates.

Intuitively, when any two explanatory variables are changing in nearly the same way, it becomes extremely difficult to establish the influence of each one regressor on Y separately. For example, assume that the consumption expenditure of an individual depends on his income and liquid assets. If over a period of time income and the liquid assets change by the same proportion, the influence on consumption of one of these variables may be erroneously attributed to the other. The effects of these variables on consumption cannot be sensibly investigated, due to their high intercorrelation.

Sources of Multicollinearity :

Multicollinearity may arise for various reasons :

First, there is a tendency of economic variables to move together over time. For example, in periods of booms and rapid economic growth the basic economic magnitudes grow, though some tend to lag behind others. Thus, income, consumption, savings, investment, prices, employment, tend to rise in periods of economic expansion and decrease in periods of recession.

Secondly, the use of lagged values of some explanatory variables as separate independent factors in the relationship : For example, in consumption function it has become customary to include among the explanatory variables past as well as the present levels of income. Naturally, the successive values of a certain variable are intercorrelated. Thus, multicollinearity is almost certain to exist in distributed lag models.

Thirdly, an over determined model : This happens when the model has more explanatory variables than number of observations.

Fourthly, the data collection method employed : For example, sampling over a limited range of the values taken by the regressors in the population.

Fifthly, model specification : For example, adding polynomial terms to a regression model, especially when the range of the X variable is small.

It should be noted that although multicollinearity is usually connected with time series, it is quite frequent in cross-section data as well. For example, in a cross section sample of manufacturing firms labour and capital inputs are almost always intercorrelated, because large firms tend to have large quantities of both factors while small firms usually have smaller quantities of both labour and capital.

However, multicollinearity tends to be more common and more serious problem in time series.

4.14. Consequences of Multicollinearity

Multicollinearity problem arises in a multiple regression model. Multicollinearity simply means the existence of correlation among the independent (explanatory) variables.

Let us consider a multiple regression model with two independent (explanatory) variables,

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, $i = 1, 2, \ldots, n$. Here, if there is correlation (product moment) between two explanatory variables X_1 and X_2, then there is the problem of multicollinearity in the model. When there is perfect correlation between X_1 and X_2 i.e., when $r_{X_1X_2}$ equals to one, then there is exact multicollinearity and when $-1 \leq r_{X_1X_2} \leq +1$, then there is near exact multicollinearity.

4.14.1. Exact Multicollinearity and its Consequences

If the intercorrelation between explanatory variables is perfect ($r_{X_iX_j} = 1$, here $r_{X_1X_2} = 1$ when there are only two explanatory variables) then

(a) the estimates of the regression coefficients are indeterminate,

and (b) the standard errors of these estimates become infinitely large.

Proof : (a) Let $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ be a three variable linear regression model.

We know that $\hat{\beta} = (X'X)^{-1}X'Y$ [See Section 3.2]

$$\text{where } \hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}_{2\times 1},\ X'X = \begin{bmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i}x_{2i} \\ \Sigma x_{1i}x_{2i} & \Sigma x_{2i}^2 \end{bmatrix}_{2\times 2},\ x_{1i} = X_{1i} - \bar{X}_1,\ x_{2i} = X_{2i} - \bar{X}_2$$

and $y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$

$$\left[\begin{aligned} &\text{Since } Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \\ &\therefore \bar{Y} = \beta_0 + \beta_1 \bar{X}_1 + \beta_2 \bar{X}_2 + \bar{u} \\ &\text{Now, } Y_i - \bar{Y} = \beta_1 (X_{1i} - \bar{X}_1) + \beta_2 (X_{2i} - \bar{X}_2) \\ &\qquad + (u_i - \bar{u}) \\ &\therefore y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + u_i \end{aligned}\right] \quad (\because \bar{u} = 0)$$

Now $\hat{\beta} = (X'X)^{-1}X'Y$ exists when $(X'X)^{-1}$ exists. i.e., when $|X'X| \neq 0$.

But here we see that

$$|X'X| = \begin{vmatrix} \Sigma x_{1i}^2 & \Sigma x_{1i}x_{2i} \\ \Sigma x_{1i}x_{2i} & \Sigma x_{2i}^2 \end{vmatrix}$$

$$= \begin{vmatrix} n\sigma_{X_1}^2 & n r_{X_1X_2}\sigma_{X_1}\sigma_{X_2} \\ n r_{X_1X_2}\sigma_{X_1}\sigma_{X_2} & n\sigma_{X_2}^2 \end{vmatrix}$$

$$= n^2\sigma_{X_1}^2 \cdot \sigma_{X_2}^2 - n^2 r_{X_1X_2}^2 \sigma_{X_1}^2 \sigma_{X_2}^2$$

$$= n^2\sigma_{X_1}^2 \sigma_{X_2}^2 \left(1 - r_{X_1X_2}^2\right)$$

$$\therefore |X'X| = n^2\sigma_{X_1}^2 \sigma_{X_2}^2 \left(1 - r_{X_1X_2}^2\right)$$

Since $\frac{1}{n}\Sigma(X_{1i} - \bar{X}_1)^2$

$$= \frac{1}{n}\Sigma x_{1i}^2 = \sigma_{X_1}^2 \ \therefore \Sigma x_{1i}^2 = n\sigma_{X_1}^2.$$

Similarly, $\Sigma x_{2i}^2 = n\sigma_{X_2}^2$

$$\text{and } r_{X_1X_2} = \frac{\text{cov}(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}}$$

$$= \frac{\frac{1}{n}\Sigma(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sigma_{X_1}\cdot\sigma_{X_2}} = \frac{\frac{1}{n}\Sigma x_{1i}x_{2i}}{\sigma_{X_1}\cdot\sigma_{X_2}}$$

$$\therefore \Sigma x_{1i}x_{2i} = n\, r_{X_1X_2}\cdot\sigma_{X_1}\sigma_{X_2}$$

If there is exact multicollinearity, $r_{X_1X_2} = \pm 1$

and hence $r_{X_1X_2}^2 = 1$ and hence $|X'X| = 0$.

So, $(X'X)^{-1}$ does not exist if $r_{X_1X_2} = \pm 1$.

Hence $\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}_{2\times 1}$ remains indeterminate if $r_{X_1X_2} = \pm 1$.

This means that OLS estimators of the parameters cannot be obtained if there is exact multicollinearity in the regression model.

Alternative Proof :

Suppose that the relation to be established is $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ and that X_1 and X_2 are related with the exact relation $X_2 = kX_1$ where k is any arbitrary constant number. Since, the formulae for the estimation of the coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are : (See section 3.2) :

$$\hat{\beta}_1 = \frac{\Sigma x_{1i}y_i \cdot \Sigma x_{2i}^2 - \Sigma x_{2i}y_i \cdot \Sigma x_{1i}x_{2i}}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i}x_{2i})^2}$$

$$\text{and } \hat{\beta}_2 = \frac{\Sigma x_{2i}y_i \cdot \Sigma x_{1i}^2 - \Sigma x_{1i}y_i \cdot \Sigma x_{1i}x_{2i}}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i}x_{2i})^2}$$

If $X_{2i} = kX_{1i}$, $\bar{X}_2 = k\bar{X}_1$

$\therefore X_{2i} - \bar{X}_2 = k(X_{1i} - \bar{X}_1)$ or, $x_{2i} = kx_{1i}$

Substituting $x_2 = kx_1$, we obtain,

$$\hat{\beta}_1 = \frac{k^2\Sigma x_{1i}y_i \cdot \Sigma x_{1i}^2 - k^2\Sigma x_{1i}y_i \cdot \Sigma x_{1i}^2}{k^2\left(\Sigma x_{1i}^2\right)^2 - k^2\left(\Sigma x_{1i}^2\right)^2} = \frac{0}{0}$$

$$\text{and } \hat{\beta}_2 = \frac{k(\Sigma x_{1i}y_i)(\Sigma x_{1i}^2) - k(\Sigma x_{1i}y_i)(\Sigma x_{1i}^2)}{k^2\left(\Sigma x_{1i}^2\right)^2 - k^2\left(\Sigma x_{1i}^2\right)^2} = \frac{0}{0}$$

Therefore the parameters are indeterminate.

Proof (b) : If $r_{X_iX_j} = 1$ (here $r_{X_1X_2} = 1$) the standard errors of estimates become infinitely large. In the regression model, $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, if X_1 and X_2 are perfectly correlated ($X_2 = kX_1$), the variances of $\hat{\beta}_1$ and $\hat{\beta}_2$ will be :

$$\text{var}(\hat{\beta}_1) = \frac{\sigma_u^2 \Sigma x_{2i}^2}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i}x_{2i})^2} \qquad \text{(See Section 3.2.2)}$$

$$\text{and } \text{var}(\hat{\beta}_2) = \frac{\sigma_u^2 \Sigma x_{1i}^2}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i}x_{2i})^2}$$

Substituting $x_2 = kx_1$ we get

$$\text{var}(\hat{\beta}_1) = \frac{\sigma_u^2 k^2 \Sigma x_{1i}^2}{k^2\Sigma x_{1i}^2 \cdot \Sigma x_{1i}^2 - k^2(\Sigma x_{1i}^2)^2} = \frac{\sigma_u^2 \Sigma x_{1i}^2}{0} = \text{undefined i.e., infinitely large.}$$

$$\text{Similarly, } \text{var}(\hat{\beta}_2) = \frac{\sigma_u^2 \Sigma x_{1i}^2}{k^2\left(\Sigma x_{1i}^2\right)^2 - k^2\left(\Sigma x_{1i}^2\right)^2} = \frac{\sigma_u^2 \Sigma x_{1i}^2}{0} = \text{undefined i.e., infinitely large.}$$

Thus the variances of the estimates and hence standard errors of estimates become infinitely large unless $\sigma_u^2 = 0$.

4.14.2. Near Exact Multicollinearity and its Consequences

In the case of near exact multicollinearity in the regression model $Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+u_i$ ($y_i=\beta_1x_{1i}+\beta_2x_{2i}+u_i$ in the case of deviation from means) we have $-1<r_{X_1X_2}<1$ and $r^2_{X_1X_2}<1$.

Hence $|X'X|=n^2\sigma^2_{X_1}\sigma^2_{X_2}\left(1-r^2_{X_1X_2}\right)>0$ i.e., $|X'X|\neq 0$

Hence $(X'X)^{-1}$ exists and $\hat{\beta}=(X'X)^{-1}X'Y$ also exists. So, in the presence of near exact multicollinearity the OLS estimators of the parameters can be estimated.

Here $|\hat{\beta}_1-\beta_1|$ is the sampling error of the statistic $\hat{\beta}_1$.

$|\hat{\beta}_2-\beta_2|$ is the sampling error of the statistic $\hat{\beta}_2$.

We know that $\hat{\beta}=(X'X)^{-1}X'Y$ where $[Y=X\beta+u]$

$=(X'X)^{-1}X'(X\beta+u)$

$=(X'X)^{-1}X'X\beta+(X'X)^{-1}X'u$

$\therefore \hat{\beta}=\beta+(X'X)^{-1}X'u$

$$\text{or, } \begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{bmatrix}=\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix}+\begin{bmatrix}\Sigma x_1^2 & \Sigma x_1x_2\\ \Sigma x_1x_2 & \Sigma x_2^2\end{bmatrix}^{-1}\begin{bmatrix}\Sigma x_1u\\ \Sigma x_2u\end{bmatrix}$$

$$\left\{\text{Here } X=\begin{bmatrix}x_{11} & x_{21}\\ x_{12} & x_{22}\\ \vdots & \vdots\\ x_{1n} & x_{2n}\end{bmatrix}_{n\times 2} \text{ and } X'=\begin{bmatrix}x_{11} & x_{12} & \cdots & x_{1n}\\ x_{21} & x_{22} & \cdots & x_{2n}\end{bmatrix}_{2\times n}\right.$$

$$u=\begin{bmatrix}u_1\\ u_2\\ \vdots\\ u_n\end{bmatrix}_{n\times 1} \quad \therefore X'u=\begin{bmatrix}x_{11} & x_{12} & \cdots & x_{1n}\\ x_{21} & x_{22} & \cdots & x_{2n}\end{bmatrix}\begin{bmatrix}u_1\\ u_2\\ \vdots\\ u_n\end{bmatrix}$$

$$\left.=\begin{bmatrix}x_{11}u_1+x_{12}u_2+\cdots+x_{1n}u_n\\ x_{21}u_1+x_{22}u_2+\cdots+x_{2n}u_n\end{bmatrix}=\begin{bmatrix}\Sigma x_1u\\ \Sigma x_2u\end{bmatrix}\right\}$$

$$\text{or, } \begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{bmatrix}=\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix}+\begin{bmatrix}\Sigma x_2^2 & -\Sigma x_1x_2\\ -\Sigma x_1x_2 & \Sigma x_1^2\end{bmatrix}\times\frac{1}{\begin{vmatrix}\Sigma x_1^2 & \Sigma x_1x_2\\ \Sigma x_1x_2 & \Sigma x_2^2\end{vmatrix}}\begin{bmatrix}\Sigma x_1u\\ \Sigma x_2u\end{bmatrix}$$

$$\text{or, } \begin{bmatrix}\hat{\beta}_1-\beta_1\\ \hat{\beta}_2-\beta_2\end{bmatrix}=\begin{bmatrix}\dfrac{\Sigma x_2^2\cdot\Sigma x_1u-\Sigma x_1x_2\cdot\Sigma x_2u}{\Sigma x_1^2\Sigma x_2^2-(\Sigma x_1x_2)^2}\\[2ex] \dfrac{\Sigma x_1^2\cdot\Sigma x_2u-\Sigma x_1x_2\cdot\Sigma x_1u}{\Sigma x_1^2\Sigma x_2^2-(\Sigma x_1x_2)^2}\end{bmatrix}$$

$$\text{or, } \begin{bmatrix}\hat{\beta}_1-\beta_1\\ \hat{\beta}_2-\beta_2\end{bmatrix}=\begin{bmatrix}\dfrac{\Sigma x_2^2\cdot\Sigma x_1u-\Sigma x_1x_2\cdot\Sigma x_2u}{n\sigma^2_{X_1}\cdot n\sigma^2_{X_1}-n^2r^2_{X_1X_2}\cdot\sigma^2_{X_1}\sigma^2_{X_2}}\\[2ex] \dfrac{\Sigma x_1^2\cdot\Sigma x_2u-\Sigma x_1x_2\cdot\Sigma x_1u}{n\sigma^2_{X_1}\cdot n\sigma^2_{X_2}-n^2r^2_{X_1X_2}\cdot\sigma^2_{X_1}\sigma^2_{X_2}}\end{bmatrix}$$

$$\hat{\beta}_1-\beta_1=\frac{\Sigma x_2^2\cdot\Sigma x_1u-\Sigma x_1x_2\Sigma x_2u}{n^2\sigma^2_{X_1}\sigma^2_{X_2}\left(1-r^2_{X_1X_2}\right)}$$

$$=\frac{n\sigma^2_{X_2}\cdot\Sigma x_1u}{n^2\sigma^2_{X_1}\sigma^2_{X_2}\left(1-r^2_{X_1X_2}\right)}-\frac{nr_{X_1X_2}\sigma_{X_1}\sigma_{X_2}\cdot\Sigma x_2u}{n^2\sigma^2_{X_1}\sigma^2_{X_2}\left(1-r^2_{X_1X_2}\right)}$$

$$=\frac{\Sigma x_1u}{n\sigma^2_{X_1}\left(1-r^2_{X_1X_2}\right)}-\frac{r_{X_1X_2}\cdot\Sigma x_2u}{n\sigma_{X_1}\sigma_{X_2}\left(1-r^2_{X_1X_2}\right)}$$

From this expression it is clear that $(\hat{\beta}_1-\beta_1)$ is the sampling error of the statistic $\hat{\beta}_1$ which depends on the correlation coefficient between X_1 and X_2.

$$\text{Similarly, } \hat{\beta}_2-\beta_2=\frac{\Sigma x_2u}{n\sigma^2_{X_2}\left(1-r^2_{X_1X_2}\right)}-\frac{r_{X_1X_2}\cdot\Sigma x_1u}{n\sigma_{X_1}\sigma_{X_2}\left(1-r^2_{X_1X_2}\right)}$$

This also shows that the sampling error of the statistic $\hat{\beta}_2$ depends on the nature and degree of the correlation coefficient between X_1 and X_2.

Note : The higher the degree of multicollinearity, the higher will be the variance of the OLS estimators of the parameters. In other words, **if $r^2_{X_1X_2}$ or $|r_{X_1X_2}|$ is high, variances of $\hat{\beta}_1$ and $\hat{\beta}_2$ are also high.** Hence the BLUE properties of the estimators no longer remain valid. This can be proved as follows :

Proof : We know that for a three variable linear regression model, variance-covariance matrix of $\hat{\beta}$ is given by, $D(\hat{\beta})=\sigma_u^2(X'X)^{-1}$ [See Section 3.2.2]

$$\text{or, } D(\hat{\beta})=\begin{bmatrix}\text{var}(\hat{\beta}_1) & \text{cov}(\hat{\beta}_1,\hat{\beta}_2)\\ \text{cov}(\hat{\beta}_1,\hat{\beta}_2) & \text{var}(\hat{\beta}_2)\end{bmatrix}=\sigma_u^2\begin{bmatrix}\Sigma x_1^2 & \Sigma x_1x_2\\ \Sigma x_1x_2 & \Sigma x_2^2\end{bmatrix}^{-1}$$

is : Estimate an equation on first differences whenever the D.W. statistics is $< R^2$. In first difference equations, we regress $(Y_t - Y_{t-1})$ on $(X_t - X_{t-1})$ (with all explanatory variables differences similarly). The implicit assumption is that the first differences of the errors $(u_t - u_{t-1})$ are independent. For instances, if $Y_t = \alpha + \beta X_t + u_t$ is the regression equation, then $Y_{t-1} = \alpha + \beta X_{t-1} + u_{t-1}$ and we have by subtraction $(Y_t - Y_{t-1}) = \beta(X_t - X_{t-1}) + (u_t - u_{t-1})$.

If the errors in this equation are independent, we can estimate the equation by OLS. However, since the constant term α disappears after subtraction, we should be estimating the regression equation with no constant term. Often, we find a constant term also included in regression equation with first differences. This procedure is valid only if there is a linear trend term in the original equation.

4.13. Multicollinearity–Meaning and Sources

Very often, the data we use in multiple regression analysis cannot give decisive answers to the questions we pose. This is because the standard errors may be very high or the 't' ratios may be very low. Confidence intervals for the parameters of interest may thus be very wide. This sort of situation occurs when explanatory variables display little variation and / or high intercorrelations. The situation where the explanatory variables are highly intercorrelated is referred to as multicollinearity. When the explanatory variables are highly intercorrelated, it becomes difficult to disentangle the separate effects of each of the explanatory variables on the explained variable.

If the explanatory variables are perfectly linearly correlated, that is, if the correlation coefficient of those variables is equal to unity (*i.e.* $r_{X_iX_j} = 1$) the parameters become indeterminate : it is impossible to obtain numerical values for each parameter separately and the method of least squares breaks down. At the other extreme if the explanatory variables are not intercorrelated at all (*i.e.*, the correlation coefficient for all these variables is equal to zero), the variables are called **orthogonal** and there are no problems concerning the estimates of the coefficients, at least so far as multicollinearity is concerned.

Actually, in the case of orthogonal X's, there is no need to perform a multiple regression analysis ; each parameter, β_i, can be estimated by a simple regression of Y on the corresponding explanatory variable X_i : $Y = f(X_i)$.

In practice, neither of the above two extreme cases (of orthogonal X's or perfect collinear X's) is often met.

In most cases there is some *degree* of intercorrelation among the explanatory variables, due to the interdependence of many economic magnitudes over time.

Multicollinearity is not a condition that either exists or does not exist in economic functions, but rather a phenomenon inherent in most relationships due to the nature of economic magnitudes. There is no conclusive evidence concerning the degree of collinearity which, if present, will affect seriously the parameter estimates.

Intuitively, when any two explanatory variables are changing in nearly the same way, it becomes extremely difficult to establish the influence of each one regressor on Y separately. For example, assume that the consumption expenditure of an individual depends on his income and liquid assets. If over a period of time income and the liquid assets change by the same proportion, the influence on consumption of one of these variables may be erroneously attributed to the other. The effects of these variables on consumption cannot be sensibly investigated, due to their high intercorrelation.

Sources of Multicollinearity :

Multicollinearity may arise for various reasons :

First, there is a tendency of economic variables to move together over time. For example, in periods of booms and rapid economic growth the basic economic magnitudes grow, though some tend to lag behind others. Thus, income, consumption, savings, investment, prices, employment, tend to rise in periods of economic expansion and decrease in periods of recession.

Secondly, the use of lagged values of some explanatory variables as separate independent factors in the relationship : For example, in consumption function it has become customary to include among the explanatory variables past as well as the present levels of income. Naturally, the successive values of a certain variable are intercorrelated. Thus, multicollinearity is almost certain to exist in distributed lag models.

Thirdly, an over determined model : This happens when the model has more explanatory variables than number of observations.

Fourthly, the data collection method employed : For example, sampling over a limited range of the values taken by the regressors in the population.

Fifthly, model specification : For example, adding polynomial terms to a regression model, especially when the range of the X variable is small.

It should be noted that although multicollinearity is usually connected with time series, it is quite frequent in cross-section data as well. For example, in a cross section sample of manufacturing firms labour and capital inputs are almost always intercorrelated, because large firms tend to have large quantities of both factors while small firms usually have smaller quantities of both labour and capital.

However, multicollinearity tends to be more common and more serious problem in time series.

4.14. Consequences of Multicollinearity

Multicollinearity problem arises in a multiple regression model. Multicollinearity simply means the existence of correlation among the independent (explanatory) variables.

Let us consider a multiple regression model with two independent (explanatory) variables,

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$, $i = 1, 2, \ldots, n$. Here, if there is correlation (product moment) between two explanatory variables X_1 and X_2, then there is the problem of multicollinearity in the model. When there is perfect correlation between X_1 and X_2, i.e., when $r_{X_1X_2}$ equals to one, then there is exact multicollinearity and when $-1 \leq r_{X_1X_2} \leq +1$, then there is near exact multicollinearity.

4.14.1. Exact Multicollinearity and its Consequences

If the intercorrelation between explanatory variables is present ($r_{X_iX_j} = 1$, here $r_{X_1X_2} = 1$ when there are only two explanatory variables) then

$$\text{VIF} = \frac{1}{\left(1 - r^2_{X_1X_2}\right)}$$

(VIF shows how the variance of an estimator is inflated by the presence of multicollinearity) As $r^2_{X_1X_2}$ approaches 1, the VIF approaches infinity. This means that as the extent of collinearity increases, the variance of an estimator increases and in the limit it can become infinite. If there is no collinearity between X_1 and X_2 (i.e., when $r^2_{X_1X_2} = 0$), VIF will be 1.

Using this definition, we can express the variances of the estimators as, $\text{var}(\hat{\beta}_1) = \frac{\sigma_u^2}{\Sigma x_{1i}^2} \cdot \text{VIF}$ and $\text{var}(\hat{\beta}_2) = \frac{\sigma_u^2}{\Sigma x_{2i}^2} \cdot \text{VIF}$ which show that the variances of $\hat{\beta}_1$ and $\hat{\beta}_2$ are directly proportional to the VIF.

This result can be easily extended to K variable model where $\text{var}(\hat{\beta}_j) = \frac{\sigma_u^2}{\Sigma x_j^2} \cdot \text{VIF}_j$, $j = 1, 2, \ldots, K-1$. From this expression we see that $\text{var}(\hat{\beta}_j)$ is proportional to σ_u^2 and VIF but inversely proportional to Σx_j^2. It should be noted that the inverse of the VIF is called **tolerance (TOL)**

i.e., $\text{TOL} = \frac{1}{\text{VIF}} = 1 - r^2_{X_1X_2}$.

It should be noted that when $r^2_{X_1X_2} = 1$, TOL = 0 and when $r^2_{X_1X_2} = 0$, TOL = 1. Because of the intimate connection between VIF and TOL, one can use them interchangably.

2. Wider confidence intervals

Because of the large standard errors, the confidence intervals for the relevant population parameters tend to be larger. This is shown in the following table :

Table 4.5.

The effect of increasing collinearity on the 95% confidence interval for β_1 :

$$\hat{\beta}_1 \pm \tau_{\alpha/2} \cdot SE(\hat{\beta}_1) = \hat{\beta}_1 \pm 1.96 \cdot SE(\hat{\beta}_1) = \hat{\beta}_1 \pm 1.96\sqrt{\frac{\sigma_u^2}{\Sigma x_{1i}^2} \cdot \text{VIF}}$$

where $\text{VIF} = \frac{1}{1 - r^2_{X_1X_2}}$, $\text{var}(\hat{\beta}_1) = \frac{\sigma_u^2}{\Sigma x_{1i}^2\left(1 - r^2_{X_1X_2}\right)}$, $SE(\hat{\beta}_1) = \sqrt{\text{var}(\hat{\beta}_1)}$

Value of $r_{X_1X_2}$	99% Confidence Interval for β_1
0.00	$\hat{\beta}_1 \pm 1.96\sqrt{\frac{\sigma_u^2}{\Sigma x_{1i}^2}}$
0.50	$\hat{\beta}_1 \pm 1.96\sqrt{1.33 \times \frac{\sigma_u^2}{\Sigma x_{1i}^2}}$
0.95	$\hat{\beta}_1 \pm 1.96\sqrt{10.26 \times \frac{\sigma_u^2}{\Sigma x_{1i}^2}}$
0.995	$\hat{\beta}_1 \pm 1.96\sqrt{100 \times \frac{\sigma_u^2}{\Sigma x_{1i}^2}}$
0.999	$\hat{\beta}_1 \pm 1.96\sqrt{500 \times \frac{\sigma_u^2}{\Sigma x_{1i}^2}}$

Note : We have used the normal distribution because σ_u^2 is assumed to be known. Hence used 1.96 for the 95% confidence factor for the normal distribution.

3. Insignificant 't' Ratios

To test the null hypothesis, say $H_0 : \beta_1 = 0$ against the alternatives $H_1 : \beta_1 \neq 0$ or $H_1 : \beta_1 > 0$, or $H_1 : \beta_1 < 0$ we use the 't' ratio, i.e., $t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$ which follows a 't' distribution with $d.f = n - 2$ (when there are two explanatory variables in the linear regression model). We have seen that in cases of high collinearity the estimated standard errors increase significantly making 't' values smaller. In such cases the null hypothesis that the relevant true population value is zero will be accepted on the basis of the test procedure.

4. A high R^2 but few significant 't' Ratios

We consider a linear multiple regression model with K explanatory variables, given by $Y_i = \beta_0 + \beta_1X_{1i} + \beta_2X_{2i} + \beta_3X_{3i} + \cdots + \beta_KX_{Ki} + u_i$. In cases of high collinearity, it is possible to find that one or more of the partial slope coefficients are individually statistically insignificant on the basis of the 't' test. Yet R^2 in such situations may be so high that on the basis of the F test we may convincingly reject the null hypothesis, $H_0 : \beta_1 = \beta_2 = \ldots.. = \beta_K = 0$. Indeed this is one of the signals of multicollinearity – insignificant 't' values but a high overall R^2 (and a significant F value).

5. Sensitivity of OLS Estimators and their standard Errors to small changes in Data

As long as multicollinearity is not perfect estimation of regression coefficient is possible but the estimates and their standard errors become very sensitive to even the slightest change in the data. We consider two separate sets of data given in two tables (Table 4.6 and 4.7).

Table 4.6. Hypothetical data on Y, X_1 and X_2

Y	X_1	X_2
1	2	4
2	0	2
3	4	12
4	6	0
5	8	16

The estimated regression equation based on the above table (Table 4.6) is :

$\hat{Y}_i = 1.1939 + 0.4463X_{1i} + 0.0030X_{2i}$

SE : (0.7737)(0.1848) (0.0851)

t : (1.5431) (2.4151) (0.0358)

$R^2 = 0.8101$, $r_{X_1X_2} = 0.5523$

$\text{cov}(\hat{\beta}_1, \hat{\beta}_2) = -0.00868$,

$d.f = (n - 3) = (5 - 3) = 2$

Table 4.7. Hypothetical data on Y, X_1 and X_2

Y	X_1	X_2
1	2	4
2	0	2
3	4	0
4	6	12
5	8	16

The estimated regression equation based on Table 4.7 is :

$\hat{Y}_i = 1.2108 + 0.4014X_{1i} + 0.0270X_{2i}$

SE :(0.7480) (0.2721) (0.1252)

t : (1.6187) (1.4752) (0.2158)

$R^2 = 0.8143$, $r_{X_1X_2} = 0.8285$

$\text{cov}(\hat{\beta}_1, \hat{\beta}_2) = -0.0282$, $d.f = 2$

As a result of a slight change in the data we see that $\hat{\beta}_1$ which was statistically significant before at 10% level of significance, is no longer significant even at that level. It is also observed that $\text{cov}(\hat{\beta}_1, \hat{\beta}_2)$ has increased from –0.00868 to –0.0282 (a more than three fold increase). All these changes may be attributable to increased multicollinearity as $r_{X_1X_2}$ has been increased from 0.5523 to 0.8285. Similarly, the standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$ have also increased between the two regressions, a usual symptom of collinearity.

4.15. Some Illustrative Examples

We can discuss some examples where the intercorrelations between the explanatory variables are high and study the consequences. Consider the model

$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$ where $y_i = Y_i - \bar{Y}$, $x_{1i} = X_{1i} - \bar{X}_1$, $x_{2i} = X_{2i} - \bar{X}_2$.

If $x_{2i} = 2x_{1i}$ we have $y_i = \beta_1 x_{1i} + \beta_2(2x_{1i}) + u_i$ or, $y_i = (\beta_1 + 2\beta_2)x_{1i} + u_i$

Thus, only $(\beta_1 + 2\beta_2)$ would be estimable. We cannot get estimates of β_1 and β_2 separately. In this case we say that there is "Perfect multicollinearity" because x_1 and x_2 are perfectly correlated (with $r^2_{X_1X_2} = 1$). In actual practice, we encounter cases where r^2 is not exactly 1 but close to 1.

As an example, we consider the case where $\Sigma x_{1i}^2 = 200$, $\Sigma x_{1i}y_i = 350$, $\Sigma x_{1i}x_{2i} = 150$, $\Sigma x_{2i}y_i = 263$ and $\Sigma x_{2i}^2 = 113$.

The OLS estimators of β_1 and β_2 would be :

$$\hat{\beta}_1 = \frac{\Sigma x_{2i}^2 \cdot \Sigma x_{1i}y_i - \Sigma x_{1i}x_{2i}\Sigma x_{2i}y_i}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i}x_{2i})^2}$$

and $$\hat{\beta}_2 = \frac{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}y_i - \Sigma x_{1i}x_{2i}\Sigma x_{1i}y_i}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i}x_{2i})^2}$$

Putting the above given values we get $\hat{\beta}_1 = 1$ and $\hat{\beta}_2 = 1$. Suppose from the data set we drop an observation and the new values become :

$\Sigma x_{1i}^2 = 199$, $\Sigma x_{1i}y_i = 347.5$, $\Sigma x_{1i}x_{2i} = 149$, $\Sigma x_{2i}y_i = 261.5$ and $\Sigma x_{2i}^2 = 112$.

Now putting the new set of values we get $\hat{\beta}_1 = -\frac{1}{2}$ and $\hat{\beta}_2 = 3$.

Thus very small changes in the variances and covariances (due to drop of observations) produce drastic changes in the estimates of the regression parameters. Here we see that the correlation coefficient between X_1 and X_2

$$= r_{X_1X_2} = \frac{\frac{1}{n}\Sigma(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sqrt{\frac{1}{n}\Sigma(X_{1i} - \bar{X}_1)^2}\sqrt{\frac{1}{n}\Sigma(X_{2i} - \bar{X}_2)^2}}$$

$$= \frac{\Sigma x_{1i}x_{2i}}{\sqrt{\Sigma x_{1i}^2}\sqrt{\Sigma x_{2i}^2}} = \frac{150}{\sqrt{200} \times \sqrt{113}} = \frac{150}{150.33} = 0.9978 \quad \therefore r_{X_1X_2} = 0.9978$$

and $r^2_{X_1X_2} = 0.995$ which is very high.

In practice, addition or deletion of observations would produce changes in the variances and covariances. Thus, one of the consequences of high correlation between X_1 and X_2 is that the parameter estimates would be very sensitive to the addition or deletion of observations.

One other symptom of the multicollinearity problem that is often mentioned is that the standard errors of the estimated regression coefficients will be very high. However, high values of $r^2_{X_1X_2}$ need not necessarily imply high standard errors and vice versa. In a three variable linear regression model (a model with two explanatory variables) we have,

$$\text{var}(\hat{\beta}_1) = \frac{\sigma_u^2}{\Sigma x_{1i}^2\left(1 - r^2_{X_1X_2}\right)} \quad \text{and} \quad \text{var}(\hat{\beta}_2) = \frac{\sigma_u^2}{\Sigma x_{2i}^2\left(1 - r^2_{X_1X_2}\right)}$$

and $$\text{cov}(\hat{\beta}_1, \hat{\beta}_2) = \frac{-\sigma_u^2 r_{X_1X_2}}{n\sigma_{X_1}\sigma_{X_2}\left(1 - r^2_{X_1X_2}\right)}$$

Thus the varance of $\hat{\beta}_1$ will be high if : (i) σ_u^2 is high, (ii) Σx_{1i}^2 is low, (iii) $r^2_{X_1X_2}$ is high. Even if $r^2_{X_1X_2}$ is high, if σ_u^2 is low and Σx_{1i}^2 is high we will not have the problem of high standard errors. On the other hand, even if $r^2_{X_1X_2}$ is low, the standard errors can be high if σ_u^2 is high and Σx_{1i}^2 is low (i.e., there is not sufficient variation in X_1).

What this suggests is that high values of $r^2_{X_1X_2}$ do not tell us anything about whether we have a multicollinearity problem or not.

Example 4.5. The following table shows time-series data for the period 2010-2019 on clothing expenditure, disposable income, liquid assets, a price index for clothing items and a general price index for a certain country.

Table 4.8

Year	Expenditure on clothing (£ m)	Disposable income (£ m)	Liquid assets (£ m)	Price index for clothing 2014 = 100	General price index 2014 = 100
2010	8.4	82.9	17.1	92	94
2011	9.6	88.0	21.3	93	96
2012	10.4	99.9	25.1	96	97
2013	11.4	105.3	29.0	94	97
2014	12.2	117.7	34.0	100	100
2015	14.2	131.0	40.0	101	101
2016	15.8	148.2	44.0	105	104
2017	17.9	161.8	49.0	112	109
2018	19.3	174.2	51.0	112	111
2019	20.8	184.7	53.0	112	111

Assuming that consumer's expenditure on clothing is influenced by all the factors included in the above table, so that the demand function for clothing is given by,

$C = \beta_0 + \beta_1 Y + \beta_2 P_c + \beta_3 L + \beta_4 P_0 + u$

where C = Expenditure on clothing
Y = Disposable income
L = Liquid assets
P_c = Price index for clothing
P_0 = General price index

By applying OLS method the following estimated regression results were obtained

$$C = -13.53 + 0.097\,Y - 0.199\,P_c + 0.015\,L + 0.34\,P_0$$

SE : (7.5) (0.03) (0.09) (0.05) (0.15)

$R^2 = 0.988$, ESS = 28.15, RSS = 0.33, TSS = 28.48

Applying analysis of variance to test the overall significance of the fit we find

$$F \text{ (observed)} = \frac{ESS / K - 1}{RSS / n - K} = \frac{28.15 / 4}{0.33 / 5} = 106.62$$

Here K = number of parameters including the constant intercept term = 5
and n = number of years (sample size) = 10

From the table value we see that $F_{0.05,\, K-1,\, n-K} = F_{0.05}$; 4, 5 = 5.19.

So, F (observed) = 106.62 > $F_{0.05;\, 4,5}$ = 5.19. and we reject the null hypothesis accepting the alternative that there is a significant relationship between clothing expenditure and the explanatory variables.

However, all the explanatory variables are seriously multicollinear as can be seen by the simple correlation coefficients :

$$r_{YL} = 0.993,\; r_{YP_c} = 0.980,\; r_{YP_0} = 0.987,\; r_{LP_c} = 0.964,\; r_{LP_0} = 0.973,\; r_{P_cP_0} = 0.991$$

To explore the effects of multicollinearity we compute the elementary regressions :

(1) $\hat{C} = \hat{a}_0 + \hat{a}_1 Y = -1.24 + 0.118Y,\ R^2 = 0.995$

SE : (0.37) (0.002)

(2) $\hat{C} = \hat{b}_0 + \hat{b}_1 P_c = -38.51 + 0.516\,P_c,\ R^2 = 0.951$

SE : (4.20) (0.04)

(3) $\hat{C} = \hat{c}_0 + \hat{c}_1 L = 2.11 + 0.327\,L,\ R^2 = 0.967$

SE : (0.81) (0.02)

(4) $\hat{C} = \hat{d}_0 + \hat{d}_1 P_0 = -53.65 + 0.663\,P_0,\ R^2 = 0.977$

SE : (3.63) (0.03)

We choose the first elementary regression [$C = f(Y)$] as the first step in our analysis as income seems to be the most important explanatory variable during the period under consideration. We then introduce the remaining explanatory variables gradually into the function :

The results are shown in the following table :

Table : 4.9

Function	$\hat{\beta}_0$ constant	$\hat{\beta}_1$ (Y)	$\hat{\beta}_2$ (P_c)	$\hat{\beta}_3$ (L)	$\hat{\beta}_4$ (P_0)	R^2
$C = f(Y)$:	−1.24 (0.37)	0.118 (0.002)	—	—	—	0.995
$C = f(Y, P_c)$:	1.40 (4.92)	0.126 (0.01)	−0.036 (0.07)	—	—	0.996
$C = f(Y, P_c, L)$:	0.94 (5.17)	0.138 (0.02)	−0.034 (0.06)	−0.037 (0.05)	—	0.996
$C = f(Y, P_c, P_0)$:	−12.76 (6.52)	0.104 (0.01)	−0.188 (0.07)	—	0.319 (0.12)	0.997
$C = f(Y, P_c, L, P_0)$:	−13.53 (7.5)	0.097 (0.03)	−0.199 (0.09)	0.015 (0.05)	0.34 (0.15)	0.988

Note : The numbers in brackets are the standard errors of the estimates.

Change in income seems to be the important factor in explaining the variation in clothing expenditure. The introduction of P_c improves slightly the overall R^2. The signs of the $\hat{\beta}$'s are correct but the standard errors show that $\hat{\beta}_2$ is not statistically significant. The high intercorrelation of Y and P_c does not affect the significance of $\hat{\beta}_1$. The introduction of liquid assets does not give a satisfactory estimate for either β_2 or β_3. Clearly the high intercorrelation of P_c and L makes it impossible to obtain separate meaningful estimates of β_2 and β_3. However, the estimate $\hat{\beta}_1$ is not affected, despite the serious intercorrelation of Y, P_c and L. Thus L may be considered as a superfluous variable.

Dropping L and introducing P_0 in the function we obtain a better overall fit. R^2 is slightly increased and all the parameter estimates have correct signs and are statistically significant. Despite the high degree of collinearity of all regressors the standard errors are not large.

The regression with all four explanatory variables shows that the effect of multicollinearity is not serious for $\hat{\beta}_1$ and $\hat{\beta}_2$. The coefficient of L, i.e., $\hat{\beta}_3$ is not significant, so that L is clearly a superfluous variable. Thus the best fit is obtained from the function, $C = f(Y, P_F, P_O)$.

4.16. Tests for Detecting Multicollinearity

Since multicollinearity is essentially a sample phenomenon, arising out of largely non- experimental data collected in most social sciences, we do not have one unique method of detecting it or measuring its strength.

We are considering some of the common measures (rules) used for detecting multicollinearity.

1. High R^2 but few significant t ratios

If R^2 is high, say, in excess of 0.8, the F test in most cases will reject the hypothesis that the partial slope coefficients are simultaneously equal to zero, but the individual 't' tests will show that none or very few of the partial slope coefficients are statistically different from zero.

2. High Pairwise correlations among regressors

Let us consider a linear regression model with two explanatory variables, given by, $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$.

Here multicollinearity problem means the strong correlation between X_1 and X_2.

In order to examine whether X_1 and X_2 are correlated or not we have to calculate the simple correlation coefficient between X_1 and X_2, denoted by $r_{X_1X_2}$.

We know that $0 \le r^2_{X_1X_2} \le 1$.

If we find that $r^2_{X_1X_2}$ is very close to zero, then the problem of multicollinearity is not so serious. But if we find that $r^2_{X_1X_2}$ is very close to 1, then the problem of multicollinearity may be considered to be serious.

Now we consider a regression model with more than two independent variables. Suppose we take four independent variables and the model takes the form :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$$

In this case we have to calculate the simple correlation coefficients between all the two possible pairs of independent variables. This means that we have to calculate the values of $r_{X_1X_2}$, $r_{X_1X_3}$, $r_{X_1X_4}$, $r_{X_2X_3}$, $r_{X_2X_4}$ and $r_{X_3X_4}$.

Now we have to examine whether the values of correlation coefficients are very close to zero or one. But the simple correlation coefficient does not give the nature and degree of the true correlation between the two independent variables considered because this is not free from the effect of the other independent variables of the model.

For example, $r_{X_1X_2}$ does not give the nature and degree of the true correlation between X_1 and X_2 because X_1 and X_2 are not free from the effects of X_3 and X_4.

In this case we have to calculate the different partial correlation coefficients :

$r_{X_1X_2 \cdot X_3X_4}$, which is the partial correlation coefficient between X_1 and X_2 eliminating the effects of X_3 and X_4.

$r_{X_1X_3 \cdot X_2X_4}$, which is the partial correlation coefficient between X_1 and X_3, eliminating the effects of X_2 and X_4.

Similarly, the values of $r_{X_1X_4 \cdot X_2X_3}$, $r_{X_2X_3 \cdot X_1X_4}$, $r_{X_2X_4 \cdot X_1X_3}$ and $r_{X_3X_4 \cdot X_1X_2}$ can be calculated. Here an indication of multicollinearity is that the partial correlation coefficients will give very high values, though the simple correlation coefficients may give very low values.

3. High value of R^2 and low values of partial correlation coefficients

We have to calculate the different partial correlation coefficients between the dependent variable Y and each independent variable, eliminating the effects of the other independent variables. This means that we have to calculate :

$r_{YX_1 \cdot X_2X_3X_4}$, which is the partial correlation coefficient between Y and X_1, eliminating the effects of X_2, X_3 and X_4 (in a linear regression model with four explanatory variables X_1, X_2, X_3 and X_4).

$r_{YX_2 \cdot X_1X_3X_4}$, which is the partial correlation coefficient between Y and X_2, eliminating the effects of X_1, X_3 and X_4.

$r_{YX_3 \cdot X_1X_2X_4}$, which is the partial correlation coefficient between Y and X_3, eliminating the effects of X_1, X_2 and X_4.

$r_{YX_4 \cdot X_1X_2X_3}$, which is the partial correlation coefficient between Y and X_4, eliminating the effects of X_1, X_2 and X_3.

Here the problem of multicollinearity is said to be serious when the multiple correlation coefficient derived from the regression line

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$, denoted by $R^2_{Y \cdot X_1X_2X_3X_4}$ takes very high value but the partial correlation coefficients $r_{YX_1 \cdot X_2X_3X_4}$, $r_{YX_2 \cdot X_1X_3X_4}$, $r_{YX_3 \cdot X_1X_2X_4}$ and $r_{YX_4 \cdot X_1X_2X_3}$ take very low values. This will happen if the independent variables X_1, X_2, X_3 and X_4 are highly intercorrelated.

4. Klein's rule for detecting multicollinearity

We consider a linear regression model with four independent variables, given by,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$$

In some cases even the partial correlation coefficients are not sufficient indicators of the seriousness of multicollinearity. Then we have to calculate the multiple correlation coefficients, regressing each independent variable on the other independent variables. This means that we have to regress X_1 on X_2, X_3 and X_4 and calculate the multiple correlation coefficient for the regression line, denoted by, $R^2_{X_1 \cdot X_2X_3X_4} = R_1^2$ (say)

Similarly, we have, $R^2_{X_2 \cdot X_1X_3X_4} = R_2^2$

$R^2_{X_3 \cdot X_1X_2X_4} = R_3^2$

and $R^2_{X_4 \cdot X_1X_2X_3} = R_4^2$

Now we have to make a comparison among these multiple correlation coefficients with the multiple correlation coefficients derived from the original model, denoted by, $R^2_{Y \cdot X_1X_2X_3X_4} = R_Y^2$ (say). By Klein's rule, multicollinearity would be regarded as a problem only if $R_Y^2 < R_i^2$ for $i = 1, 2, 3, 4$. But the multicollinearity problem is not serious if $R_Y^2 > R_i^2$ for $i = 1,2,3,4$.

In particular, if it is found that $R_Y^2 < R_2^2$, then the second variable X_2 is considered to be serious from the view point of multicollinearity problem. Similarly, if $R_Y^2 < R_3^2$, then the third independent variable X_3 may be considered to be serious from the view point of multicollinearity problem and so on.

5. The test suggested by FARRAR and GLAUBER

Let us suppose that there are four independent variables X_1, X_2, X_3 and X_4 in the regression model. Let us construct the following correlation matrix,

$$R_X = \begin{bmatrix} 1 & r_{12} & r_{13} & r_{14} \\ r_{21} & 1 & r_{23} & r_{24} \\ r_{31} & r_{32} & 1 & r_{34} \\ r_{41} & r_{42} & r_{43} & 1 \end{bmatrix} \quad \begin{aligned} \text{where } & r_{11} = r_{22} = r_{33} = r_{44} = 1 \\ \text{and } & r_{12} = r_{21} \\ & r_{13} = r_{31} \\ & r_{23} = r_{32} \\ & r_{24} = r_{42} \end{aligned}$$

In the case of two explanatory variables X_1 and X_2 we have,

$$R_X = \begin{bmatrix} 1 & r_{12} \\ r_{21} & 1 \end{bmatrix} \quad \begin{aligned} \text{where } & r_{11} = r_{22} = 1 \\ \text{and } & r_{12} = r_{21} \end{aligned}$$

and for three explanatory variables X_1, X_2 and X_3 we have,

$$R_X = \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix} \quad \begin{aligned} \text{where } & r_{11} = r_{22} = r_{33} = 1 \\ \text{and } & r_{12} = r_{21} \\ & r_{23} = r_{32} \\ & r_{31} = r_{13} \end{aligned}$$

Now if it is found that the determinant of R_X i.e. $|R_X| = 1$, there is no problem of multicollinearity. If $|R_x| = 0$, there is the problem of exact multicollinearity and if $|R_x|$ is very close to zero, then there is near exact multicollinearity but serious. *FARRAR* and *GLAUBER* suggested in testing the null hypothesis $H_0 : |R_X| = 0$. But this test is not very meaningful because multicollinearity is the problem with the sample we have, not with the population.

4.17. Solutions to the Problem of Multicollinearity

There are six commonly used methods to solve the problem of multicollinearity :

(1) Dropping of variables

(2) Using extraneous estimates

(3) Ridge Regression

(4) Using ratios of first differences

(5) Using principal components

(6) Getting more data.

We can now briefly explain these methods as follows :

(1) Dropping of variables :

Let us consider a linear regression model with two independent variables, given by, $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ and we are interested in estimating β_1 but the problem is that X_1 and X_2 are highly correlated.

Now taking deviations from means we get

$Y_i - \bar{Y} = \beta_1(X_{1i} - \bar{X}_1) + \beta_2(X_{2i} - \bar{X}_2) + (u_i - \bar{u})$ [where $\bar{Y} = \beta_0 + \beta_1\bar{X}_1 + \beta_2\bar{X}_2 + \bar{u}$]

or, $y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$ .... (1) $\because \bar{u} = 0$

where $y_i = Y_i - \bar{Y}$; $x_{1i} = X_{1i} - \bar{X}_1$ and $x_{2i} = X_{2i} - \bar{X}_2$. Here u_i satisfies all the CLRM properties. We can estimate β_1 and β_2 by OLS method.

Let the OLS estimators of β_1 and β_2 be $\hat{\beta}_1$ and $\hat{\beta}_2$ where

$$\hat{\beta}_1 = \frac{\Sigma x_{2i}^2 \cdot \Sigma x_{1i} y_i - \Sigma x_{1i} x_{2i} \cdot \Sigma x_{2i} y_i}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2} \text{ and } \operatorname{var}(\hat{\beta}_1) = \frac{\sigma_u^2}{\Sigma x_{1i}^2 \left(1 - r^2_{X_1 X_2}\right)}$$

Since we are interested in estimating β_1, in order to avoid the problem of multicollinearity we can drop X_2 and specify the model as follows :

$y_i = \beta_1 x_{1i} + u_i$ .... (2)

Now OLS estimator of β_1 obtained from model (2) is denoted by β_1^* where

$$\beta_1^* = \frac{\Sigma x_{1i} y_i}{\Sigma x_{1i}^2}$$

$$\therefore \beta_1^* = \frac{\Sigma x_{1i}(\beta_1 x_{1i} + \beta_2 x_{2i} + u_i)}{\Sigma x_{1i}^2} \text{ [from (1), } y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + u_i]$$

$$\text{or, } \beta_1^* = \beta_1 \frac{\Sigma x_{1i}^2}{\Sigma x_{1i}^2} + \beta_2 \frac{\Sigma x_{1i} x_{2i}}{\Sigma x_{1i}^2} + \frac{\Sigma x_{1i} u_i}{\Sigma x_{1i}^2}$$

$$\text{or, } \beta_1^* = \beta_1 + \beta_2 \cdot \frac{\Sigma x_{1i} x_{2i}}{\Sigma x_{1i}^2} + \frac{\Sigma x_{1i} u_i}{\Sigma x_{1i}^2}$$

$$\text{Now, } E(\beta_1^*) = \beta_1 + \beta_2 \frac{\Sigma x_{1i} x_{2i}}{\Sigma x_{1i}^2} + \frac{\Sigma x_{1i} E(u_i)}{\Sigma x_{1i}^2}$$

$$\text{or, } E(\beta_1^*) = \beta_1 + \beta_2 \frac{\Sigma x_{1i} x_{2i}}{\Sigma x_{1i}^2}, \text{ where } E(u_i) = 0$$

$$\therefore E(\beta_1^*) - \beta_1 = \beta_2 \frac{\Sigma x_{1i} x_{2i}}{\Sigma x_{1i}^2} = \text{Bias}$$

$$\text{Now } E(\beta_1^*) - \beta_1 = \beta_2 \cdot \frac{n r_{X_1 X_2} \cdot \sigma_{X_1} \sigma_{X_2}}{n \sigma^2_{X_1}} = \beta_2 \cdot \frac{r_{X_1 X_2} \cdot \sigma_{X_2}}{\sigma_{X_1}}$$

If the two explanatory variables X_1 and X_2 are correlated, then $r_{X_1X_2} \neq 0$ and hence $E(\beta_1^*) \neq \beta_1$. Hence β_1^* which is the OLS estimator of β_1 obtained from model (2) is not an unbiased estimator of β_1. This bias arises due to the dropping of one explanatory variable which leads to misspecification of the relationship. This is why this bias is called the **Specification bias.**

Mean Square Error (MSE) :

Instead of var (β_1^*) we calculate MSE (β_1^*) where $E(\beta_1^*) \neq \beta_1$ i.e. β_1^* is a biased estimator of β_1.

Now MSE $(\beta_1^*) = E[\beta_1^* - \beta_1]^2$

$$= E\left[\beta_2 \cdot r_{X_1X_2} \cdot \frac{\sigma_{X_2}}{\sigma_{X_1}} + \frac{\Sigma x_{1i}u_i}{\Sigma x_{1i}^2}\right]^2$$

$$\left[\text{Since } \beta_1^* = \beta_1 + \beta_2\frac{\Sigma x_{1i}x_{2i}}{\Sigma x_{1i}^2} + \frac{\Sigma x_{1i}u_i}{\Sigma x_{1i}^2}\right.$$

$$\text{or, } \beta_1^* - \beta_1 = \beta_2\frac{\Sigma x_{1i}x_{2i}}{\Sigma x_{1i}^2} + \frac{\Sigma x_{1i}u_i}{\Sigma x_{1i}^2} = \beta_2\frac{n\, r_{X_1X_2}\sigma_{X_1}\sigma_{X_2}}{n\sigma_{X_1}^2} + \frac{\Sigma x_{1i}u_i}{\Sigma x_{1i}^2}$$

$$\left.\therefore \beta_1^* - \beta_1 = \beta_2 \cdot r_{X_1X_2}\frac{\sigma_{X_2}}{\sigma_{X_1}} + \frac{\Sigma x_{1i}u_i}{\Sigma x_{1i}^2}\right]$$

$$\therefore \text{MSE}(\beta_1^*) = E\left[\beta_2 r_{X_1X_2}\frac{\sigma_{X_2}}{\sigma_{X_1}} + \frac{\Sigma x_{1i}u_i}{\Sigma x_{1i}^2}\right]^2$$

$$= E\left[\beta_2^2 \cdot r_{12}^2\frac{\sigma_{X_2}^2}{\sigma_{X_1}^2} + 2\beta_2 r_{12}\frac{\sigma_{X_2}}{\sigma_{X_1}}\cdot\frac{\Sigma x_{1i}u_i}{\Sigma x_{1i}^2} + \left(\frac{\Sigma x_{1i}u_i}{\Sigma x_{1i}^2}\right)^2\right] \text{ [where we put } r_{X_1X_2} = r_{12}]$$

$$= \left[\beta_2^2 \cdot r_{12}^2 \cdot \frac{\sigma_{X_2}^2}{\sigma_{X_1}^2}\right] + 2\beta_2 r_{12}\frac{\sigma_{X_2}}{\sigma_{X_1}}\cdot\frac{\Sigma x_{1i}E(u_i)}{\Sigma x_{1i}^2} + E\left[\frac{\Sigma x_{1i}u_i}{\Sigma x_{1i}^2}\right]^2$$

$$= \beta_2^2 \cdot r_{12}^2 \cdot \frac{\sigma_{X_2}^2}{\sigma_{X_1}^2} + E\left[\frac{\Sigma x_{1i}u_i}{\Sigma x_{1i}^2}\right]^2 \text{ as } E(u_i) = 0$$

$$= \beta_2^2 \cdot r_{12}^2 \cdot \frac{\sigma_{X_2}^2}{\sigma_{X_1}^2} + \frac{\sigma_u^2\Sigma x_{1i}^2}{(\Sigma x_{1i}^2)^2}, \text{ as } E(u_i^2) = \sigma_u^2.$$

$$\therefore \text{MSE}(\beta_1^*) = \beta_2^2 r_{12}^2\frac{\sigma_{X_2}^2}{\sigma_{X_1}^2} + \frac{\sigma_u^2}{\Sigma x_{1i}^2}$$

$$= \beta_2^2 r_{12}^2\frac{\Sigma x_{2i}^2}{\Sigma x_{1i}^2} + \frac{\sigma_u^2}{\Sigma x_{1i}^2} = \frac{1}{\Sigma x_{1i}^2}\left[\beta_2^2 \cdot r_{12}^2\Sigma x_{2i}^2 + \sigma_u^2\right]$$

Again we know that $\text{var}(\hat\beta_1) = \dfrac{\sigma_u^2}{\Sigma x_{1i}^2(1 - r_{12}^2)}$

$$\text{Now } \frac{\text{MSE}(\beta_1^*)}{\text{var}(\hat\beta_1)} = \frac{\dfrac{1}{\Sigma x_{1i}^2}\left[\beta_2^2\Sigma x_{2i}^2 \cdot r_{12}^2 + \sigma_u^2\right]}{\dfrac{\sigma_u^2}{\Sigma x_{1i}^2(1 - r_{12}^2)}}$$

$$= \frac{\beta_2^2 \cdot \Sigma x_{2i}^2 \cdot r_{12}^2(1 - r_{12}^2)}{\sigma_u^2} + (1 - r_{12}^2) = \left\{\frac{\beta_2^2 \cdot r_{12}^2}{\sigma_u^2/\Sigma x_{2i}^2(1 - r_{12}^2)}\right\} - r_{12}^2 + 1$$

$$= \frac{\beta_2^2 r_{12}^2}{\text{var}(\hat\beta_2)} - r_{12}^2 + 1, \text{ since } \text{var}(\hat\beta_2) = \frac{\sigma_u^2}{\Sigma x_{2i}^2(1 - r_{12}^2)}$$

$$= r_{12}^2\left[\left(\frac{\beta_2}{\sqrt{\text{var}(\hat\beta_2)}}\right)^2 - 1\right] + 1 \quad \therefore \frac{\text{MSE}(\beta_1^*)}{\text{var}(\hat\beta_1)} = r_{12}^2\left[\left(\frac{\beta_2}{\sqrt{\text{var}(\hat\beta_2)}}\right)^2 - 1\right] + 1$$

Let us put $t_2 = \dfrac{\beta_2}{\sqrt{\text{var}(\hat\beta_2)}} = \dfrac{\beta_2}{\sqrt{\dfrac{\sigma_u^2}{\Sigma x_{2i}(1 - r_{12}^2)}}}$

Here β_2 and σ_u^2 are unknown and hence t_2 is also unknown. So, t_2 is to be replaced by the suitable estimator of t_2.

We know that $\hat\beta_2$ is an unbiased estimator of β_2 and $\dfrac{\Sigma e_i^2}{n-3}$ is an unbiased estimator of σ_u^2

and $\hat t_2 = \dfrac{\hat\beta_2}{\sqrt{\dfrac{\Sigma e_i^2/(n-3)}{\Sigma x_{2i}^2(1 - r_{12}^2)}}}$ may be taken as the appropriate estimator of t_2.

Here $\dfrac{\text{MSE}(\beta_1^*)}{\text{var}(\hat\beta_1)} \gtreqless 1$ according as $\hat t_2 \gtreqless 1$

Since t_2 is unknown, we have to compute the value of $\hat t_2$.

If $|\hat t_2| > 1$, then MSE $(\beta_1^*) > \text{var}(\hat\beta_1)$

and if $|\hat t_2| < 1$, then MSE $(\beta_1^*) < \text{var}(\hat\beta_1)$

BANCROFT suggested conditional omitted variable (COV) estimator of β_1, defined as

$$\tilde{\beta}_1 = \begin{cases} \hat{\beta}_1, \text{ the OLS estimator of } \beta_1 \text{ if } |\hat{t}_2| \geq 1 \\ \beta_1^*, \text{ the COV estimator of } \beta_1 \text{ if } |\hat{t}_2| < 1 \end{cases}$$

HUNTS BERGER suggested weighted Average Estimator (WTD) of β_1

Here, instead of using $\hat{\beta}_1$ and β_1^*, depending on $\hat{t}_2$ we consider a linear combination of both, called WTD estimator of β_1, denoted by $\hat{\hat{\beta}}_1$

where $\hat{\hat{\beta}}_1 = \lambda\hat{\beta}_1 + (1-\lambda)\beta_1^*$

Two important properties of WTD estimator of β_1 are :

(i) WTD estimator of β_1 denoted by $\hat{\hat{\beta}}_1$ is a biased estimator of β_1 i.e., $E\left(\hat{\hat{\beta}}_1\right) \neq \beta_1$

and (ii) MSE $(\hat{\hat{\beta}}_1)$ is minimum if $\lambda = \dfrac{t_2^2}{1+t_2^2}$ where $t_2 = \dfrac{\beta_2}{\sqrt{\operatorname{var}(\hat{\beta}_2)}}$

FELDSTEIN studied the mean-squared error of WTD and COV estimators for different values of t_2 and $\hat{t}_2$. He argues that

(i) Omitting a collinear nuisance variable on the basis of its sample t statistic $\hat{t}_2$ is generally not advisable. OLS is preferable to any COV-estimator unless one has a strong prior notion that $|t_2|$ is <1.

(ii) The WTD estimator is generally better than COV estimator.

(iii) The WTD estimator is superior to OLS for $|t_2| \leq 1.25$ and only slightly inferior for $1.5 \leq |t_2| \leq 3.0$.

(iv) The inadequacy of collinear data should not be disguised by reporting results from the omitted variable regressions.

Even if a WTD estimator is used, one should report the OLS estimates and their standard errors to let the readers judge the extent of multicollinearity.

(ii) Use of Extraneous Information :

This method was followed in early demand studies. It was found in time series data, that income and price were both highly correlated. Hence, neither the price nor the income elasticity of demand could be estimated properly. It is often suggested in that case to estimate the income elasticity of demand, using cross section data first. Then on the basis of that estimated income elasticity of demand from cross section data we can estimate the price elasticity of demand from the time series data.

Let log $Y = \alpha + \beta$ log X_1 be an equation where Y stands for the level of demand and X_1 stands for the level of income of the consumer. Now at the same point of time we can get information on income and demand from different consumers. Regressing log Y on log X_1, taking each consumer as the unit we can estimate α and β by OLS method. Let $\hat{\beta}$ be the OLS estimator of β. Then $\hat{\beta}$ is the estimated income elasticity of demand from the cross section data.

Now we want to estimate the price elasticity of demand, using time series data. In this case we consider the model, log $Y = \beta_0 + \beta_1$ log $X_1 + \beta_2$ log X_2 where Y is the level of demand, X_1 is the level of income and X_2 is the price.

We collect the data on Y, X_1 and X_2 for different time periods and using this time series data we estimate β_0 and β_2 but not β_1. We use $\hat{\beta}$, estimated from the previous cross section study as the extraneous estimate of β_1.

Let us put $\beta_1 = \hat{\beta}$ and get

$$\log Y = \beta_0 + \hat{\beta} \log X_1 + \beta_2 \log X_2$$

$$\Rightarrow \log Y - \hat{\beta} \log X_1 = \beta_0 + \beta_2 \log X_2$$

$$\Rightarrow Z = \beta_0 + \beta_2 \log X_2 \text{ where } Z = \log Y - \hat{\beta} \log X_1.$$

Since $\hat{\beta}$ is known we can get the value of Z for each unit. Now regressing Z on log X_2 we can estimate β_0 and β_2 where OLS estimator of β_2, denoted by $\hat{\beta}_2$ is the estimated price elasticity of demand from the time series data on the basis of already estimated income elasticity of demand from the cross-section data.

There are two main problems with this procedure :

Firstly, The fact that β_1 has been estimated should be taken into account in computing the variances of the estimates of β_0 and β_2. But in most practical applications that is not done.

Secondly, The cross section estimate of the coefficient may stand for which is completely different from what the time series estimate is supposed to measure.

(iii) Ridge Regression :

One of the solutions often suggested for the multicollinearity problem is to use what is known as "ridge regression", first introduced by **HOERL** and **KENNARD**. Simply stated, the idea is to add a constant λ to the variances of the explanatory variables before solving the normal equations.

The addition of λ to the variances produces biased estimators but the argument is that if the variance can be decreased, the mean squared error (MSE) will decline. HOERL and KENNARD show that there always exists a constant $\lambda > 0$ such that

$$\sum_{i=1}^{K} \text{MSE}(\hat{\hat{\beta}}_i) < \sum_{i=1}^{K} \text{MSE}(\hat{\beta}_i)$$

where the $\hat{\hat{\beta}}_i$ are the estimators of β_i from the ridge regression, the $\hat{\beta}_i$ are the least squares estimators and K is the number of regressors.

Unfortunately, λ is a function of the regression parameters β_i and the error variance σ_u^2, which are unknown. HOERL and KENNARD suggest trying different values of λ and picking the value of λ so that the system will stabilise or the "coefficients do not have unreasonable values". Some others have suggested obtaining initial estimates of β_i and σ_u^2 and then use the estimated λ. This procedure can be iterated and we get the iterated ridge estimator. One other problem about ridge regression is the fact that it is not invariant to the units of measurement of the explanatory variables and to linear transformations of variables. Because of the deficiencies of ridge regression the method is not recommended as a general solution to the multicollinearity problem.

(iv) Using Ratios of first differences :

Let us consider the following regression model : $Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t$ where X_{1t} and X_{2t} are highly correlated. We can either solve or at least minimise the problem of multicollinearity taking ratios of first differences.

Using ratios : Let us suppose we are interested in estimating β_2, not β_1. Then dividing both sides of the regression equation by X_{1t} we get,

$$\frac{Y_t}{X_{1t}} = \frac{\beta_0}{X_{1t}} + \beta_1 + \beta_2 \frac{X_{2t}}{X_{1t}} + \frac{u_t}{X_{1t}}$$

Now if we regress $\frac{Y_t}{X_{1t}}$ on $\frac{1}{X_{1t}}$ and $\frac{X_{2t}}{X_{1t}}$, it may be that the degree of correlation between $\frac{1}{X_{1t}}$ and $\frac{X_{2t}}{X_{1t}}$ may be lower than the degree of correlation between X_{1t} and X_{2t} in this model. Let us suppose that $u_t^* = \frac{u_t}{X_{1t}}$, then

$$\text{var}\ (u_t^*) = E\left[\frac{u_t}{X_{1t}}\right]^2 = \frac{E(u_t^2)}{X_{1t}^2} = \frac{\sigma_u^2}{X_{1t}^2}$$

So, the variance of the disturbance term is no longer constant, it varies inversely with one independent variable X_{1t}. This is the problem of *heteroscedasticity.*

Using first differences :

For period '*t*' we have $Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t$ and for period $t - 1$ we have $Y_{t-1} = \beta_0 + \beta_1 X_{1,t-1} + \beta_2 X_{2,t-1} + u_{t-1}$. It we take the first differences then we get

$$Y_t - Y_{t-1} = \beta_1(X_{1t} - X_{1,t-1}) + \beta_2(X_{2t} - X_{2,t-1}) + (u_t - u_{t-1})$$

or, $\Delta Y_t = \beta_1 \Delta X_{1t} + \beta_2 \Delta X_2 t + \Delta u_t$

Now we regress ΔY_t on ΔX_{1t} and $\Delta X_2 t$. It may be that the degree of correlation between ΔX_{1t} and ΔX_{2t} may be lower than the degree of correlation between X_{1t} and X_{2t}.

The CLRM properties are :

$E(u_t) = 0,\ E(u_t^2) = \sigma_u^2,\ \ E(u_t, u_{t-1}) = 0,$

$$\begin{aligned} E(\Delta u_t^2) &= E[u_t - u_{t-1}]^2 \\ &= E[u_t^2 - 2u_t u_{t-1} + u_{t-1}^2] \\ &= E(u_t^2) - 2E(u_t, u_{t-1}) + E(u_{t-1}^2) \\ &= \sigma_u^2 - 2 \times 0 + \sigma_u^2 = 2\sigma_u^2 \end{aligned}$$

$$\begin{aligned} E[\Delta u_t, \Delta u_{t-1}] &= E[(u_t - u_{t-1})(u_{t-1} - u_{t-2})] \\ &= E[u_t u_{t-1} - u_t u_{t-2} - u_{t-1}^2 + u_{t-1} u_{t-2}] \\ &= E[u_t u_{t-1}] - E[u_t u_{t-2}] - E(u_{t-1})^2 + E[u_{t-1} u_{t-2}] \\ &= 0 - 0 - E(u_{t-1}^2) + 0 = -E(u_{t-1}^2) = -\sigma_u^2. \end{aligned}$$

when we take the first differences of the dependent variable, properties of CLRM are not satisfied. Covariances between two disturbance terms are not zero i.e., the disturbance terms are not independent of each other. This is called the problem of autocorrelation.

(v) Using Principal Components :

Another solution suggested for the multicollinearity problem is the principal component regression. Suppose we have K explanatory variables in the regression model. Then we can consider linear functions of these variables.

$$\begin{aligned} Z_1 &= a_1 X_1 + a_2 X_2 + \ldots\ldots + a_K X_K \\ Z_2 &= b_1 X_1 + b_2 X_2 + \ldots\ldots + b_K X_K \end{aligned}$$

and so on.

Suppose we choose the a's so that the variance of Z_1 is minimised subject to the condition that $a_1^2 + a_2^2 + \ldots + a_K^2 = 1$.

This is called the normalization condition. (It is required or else the variance of Z_1 can be increased indefinitely). Z_1 is then said to be the first principal component. It is a linear function of the X's that has the highest variance (subject to the normalization rule).

In this way the process of maximizing the variance of the linear function Z, subject to the condition that the sum of squares of the coefficients of the X's is equal to 1, produces K solutions. Corresponding to these, we construct K linear functions $Z_1, Z_2, \ldots, Z_K$. These are called the principal components of the X s. They can be ordered so that

$$\text{var}\ (Z_1) > \text{var}\ (Z_2) > \ldots. > \text{var}\ (Z_K).$$

Z_1, the one with highest variance, is called the first principal component, Z_2, with the next highest variance, is called the second principal component, and so on. These principal components have the following properties :

1. $\text{var}\ (Z_1) + \text{var}\ (Z_2) + \ldots + \text{var}\ (Z_K) = \text{var}\ (X_1) + \text{var}\ (X_2) + \ldots + \text{var}\ (X_K)$

2. Unlike the X s, which are uncorrelated, the Z s are orthogonal or uncorrelated. Thus these is zero multicollinearity among the Z s.

Sometimes it is suggested, instead of regressing Y on $X_1, X_2, \ldots, X_K$, we should regress Y on $Z_1, Z_2, \ldots, Z_K$. But this is not a solution to the multicollinearity problem. If we regress Y on Z s and then substitute the values of Z s in terms of X s, we finally get the same answers as before. The fact that the Z s are uncorrelated does not mean that we will get better estimates of the coefficients in the original regression equation. So, there is a point in using the principal components only if we regress Y on a subset of the Z s.

However, there are some problems with this procedure.

1. The first principal component Z_1, which has the highest variance, need not be the one that is mostly correlated with Y. In fact, there is no necessary relationship between the order of the principal components and the degree of correlation with the dependent variable Y.
2. One can think of choosing only those principal components that have high correlation with Y and discard the rest.
3. The linear combination of Z s often do not have economic meaning.
4. Changing the units of measurement of the X s will change the principal components. However, this problem can be avoided if all variables are standardised to have unit variance. However, there are some uses for the principal component method in exploratory stages of the investigation.

(v) Getting more data :

It has been suggested that multicollinearity may be avoided or reduced if we increase the size of the sample by gathering more observations. **CHRIST** says that by increasing the sample, high covariances among estimated parameters resulting from multicollinearity in an equation can be reduced, because the covariances are inversely proportional to sample size. This is true if multicollinearity is due to errors of measurement, as well as when intercorrelation happens to exist only in our original sample but not in the population of the X's.

EXERCISE

1. What is meant by autocorrelation ? How does it differ from simple correlation ? what assumptions of the CLRM will not hold if the problem of autocorrelation exists in the model ?
2. What is the meaning of the term heteroscedasticity ? How does it differ from homoscedasticity ? What assumptions of the CLRM will not hold if the problem of heteroscedasticity exists in the model ?
3. By using matrix algebra how can you identify the problem of autocorrelation as well as the problem of heteroscedasticity ?
4. Explain the possible consequences of the problem of heteroscedasticity.
5. Show that the OLS estimators will be unbiased even if there is a problem of heteroscedasticity in the CLRM.
6. Show that in presence of heteroscedasticity the OLS estimators will be unbiased but BLUE property may not be satisfied.
7. How can you estimate the regression parameters in presence of heteroscedasticity ?
8. Explain the WLS method in estimating a regression model in the presence of the problem of heteroscedasticity.
9. Explain briefly the different testing procedures used for testing the problem of heteroscedasticity.
10. Given : $Y_i = \alpha + \beta X_i + u_i$ with $E(u_i^2) = k^2 X_i^2$ prove that OLS estimates ($\hat{\alpha}$ and $\hat{\beta}$) possess greater variance than the OLS estimates of the transformed version of the model.

[**Hints :** The model is $Y_i = \alpha + \beta X_i + u_i$ where u_i satisfies all the assumptions of CLRM except that u_i is heteroscedastic, *i.e.*, $u_i \sim N(0, \sigma_{u_i}^2)$. Here the heteroscedasticity is of the form :

$E(u_i^2) = \sigma_{u_i}^2 = k^2 X_i^2$, where k is some finite constant to be estimated from the model.

Given that $\sigma_{u_i}^2 = k^2 X_i^2$ or, $k^2 = \dfrac{\sigma_{u_i}^2}{X_i^2}$ $\therefore k = \dfrac{\sigma_{u_i}}{X_i}$.

In other words, the required transformation should be $\dfrac{Y_i}{X_i} = \dfrac{\alpha}{X_i} + \dfrac{\beta X_i}{X_i} + \dfrac{u_i}{X_i}$.

or, $\dfrac{Y_i}{X_i} = \dfrac{\alpha}{X_i} + \beta + \dfrac{u_i}{X_i}$.

Now, $\text{var}\left(\dfrac{u_i}{X_i}\right) = E\left(\dfrac{u_i}{X_i}\right)^2 = \dfrac{1}{X_i^2} \cdot E(u_i^2) = \dfrac{\sigma_{u_i}^2}{X_i^2}$.

Since we have assumed $\sigma_{u_i}^2 = k^2 X_i^2$,

$\therefore \text{var}\left(\dfrac{u_i}{X_i}\right) = \dfrac{\sigma_{u_i}^2}{X_i^2} = k^2$ (a constant) which proves that new random term in the model has a finite constant variance ($= k^2$). We can therefore, apply OLS to the transformed version of the model :

$$\frac{Y_i}{X_i} = \frac{\alpha}{X_i} + \beta + \frac{u_i}{X_i}.$$

In this model we have, $\hat{\hat{\beta}} = \left(\dfrac{\bar{Y}_i}{X_i}\right) - \hat{\hat{\alpha}}\left(\dfrac{1}{X_i}\right)$

$$\text{var}\left(\hat{\hat{\beta}}\right) = \frac{k^2\Sigma\left(\frac{1}{X_i^2}\right)}{n\Sigma\left(\frac{1}{X_i^2}\right) - \left[\Sigma\frac{1}{X_i}\right]^2}$$

$$\left[\therefore \text{var}(\hat{\beta}) \text{ in OLS} = \sigma_u^2\left(\frac{\Sigma X_i^2}{n\Sigma x_i^2}\right)\right]$$

Clearly, $\text{var}(\hat{\beta}) > \text{var}(\hat{\hat{\beta}})$.

Here $\hat{\hat{\beta}}$ and $\hat{\hat{\alpha}}$ are the OLS estimators of the transformed model].

11. Consider the model $Y_i = \alpha + \beta X_i + u_i$, $u_t = \rho u_{t-1} + \epsilon_t$

    Suppose that we calculate $\hat{\beta}$ by OLS formula : $\hat{\beta} = \dfrac{\Sigma x_i y_i}{\Sigma x_i^2}$. Show that because of the autocorrelation, our usual formula for the variance of $\hat{\beta}$, namley, $\text{var}(\hat{\beta}) = \dfrac{\sigma_u^2}{\Sigma x_i^2}$ no longer holds.

12. Explain Durbin-Watson test in brief for testing the problem of autocorrelation. Point out the major limitations of Durbin-Watson test.

13. Find out the mean, variance and covariance of the autocorrelated disturbance variable when the value of the error term is autocorrelated with a first order autoregressive scheme.

14. Explain the method by means of which you can estimate the regression parameters in the presence of the problem of autocorrelation.

15. Define the term "multicollinearity". Explain how you would detect its presence in a multiple regression equation you have estimated ? What are the consequences of multicollinearity, and what are the solutions ?

16. Explain the following methods :
    (a) Ridge regression
    (b) Omitted variable regression
    (c) Principal component regression.
    What are the problems that these methods are supposed to solve ?

17. Examine whether the following statements are true or false.
    (a) In a multiple regression, a high correlation in the sample among the regressors (multicollinearity) implies that the least squares estimators of the coefficients are biased.
    (b) Whether multicollinearity is a problem or not cannot be decided by just looking at the intercorrelations between the explanatory variables.
    (c) If the coefficient estimates in an equation have high standard errors, this is evidence of high multicollinearity.
    (d) The relevant question to ask if there is high multicollinearity is not what variables to drop but what other information will help.

18. What is meant by the term "multicollinearity" ? What are the different sources of multicollinearity ?

19. What is meant by exact multicollinearity ? Examine its consequence in terms of linear regression model with two explanatory variables.

20. What is meant by near exact multicollinearity ? How does it differ from exact multicollinearity ? Examine the consequences of near exact multicollinearity in terms of a linear regression model with two explanatory variables.

21. The higher the degree of multicollinearity, the higher will be the variance of the OLS estimators of the parameters. Examine it with the help of a linear regression model with two explanatory variables.

22. What is VIF ? What is TOL ? What is the relation between the two ? What would be the value of TOL when $r^2_{X_1X_2} = 1$ and what would be the value of TOL when $r^2_{X_1X_2} = 0$ (assuming a linear regression model with two explanatory variables only) ?

23. What is meant by multicollinearity ? What are the different methods used for detecting the multicollinearity in a regression model ?

24. What is meant by multicollinearity ? Explain the following methods for detecting multicollinearity :
    (a) Klein's rule for detecting multicollinearity.
    (b) The test suggested by FARRAR and GLAUBER.

25. State and briefly explain the methods used for solving the problem of multicollinearity.

26. What is weighted average estimator (WTD) ? State its properties.

27. Explain briefly the following methods for solving the problem of multicollinearity :
    (a) Dropping of variables
    (b) Using extraneous estimates
    (c) Ridge regression
    (d) Using ratios of first differences
    (e) Using principal components
    (f) Getting more data.

28. What is specification bias ? How does it arise in the case of dropping of variable(s) from the regression model ?

29. Suppose we have the following model and initial conditions : $Y_t = 20 + X_t + u_t$, $u_t = u_{t-1} + \epsilon_t$, $u_{-1} = 5$, $u_0 = 6$, and $\epsilon_t = 0$ for all observations.
    Assuming that X_t takes the values from 1 to 20,
    (i) generate an artificial sample of 20 observations.
    (ii) Use this general sample to estimate $\hat{\alpha}$ and $\hat{\beta}$ by OLS and compute d-statistic to test for first order autocorrelation, and (iii) estimate ρ and apply OLS to the transformed data.
    Are the new estimates $\hat{\alpha}^*$ and $\hat{\beta}^*$ improved by the transformation undertaken by you?

30. The data of the following table are the OLS residuals of a consumption function : $\hat{C}_t = 3.02 + 0.93 Y_t$. Show that d-statistic = 1.42. Can you test for the existence of autocorrelation ?

Years :	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
e_t :	0.6	1.9	–1.8	–2.7	–2.9	1.4	3.3	0.3	0.8	2.3	–1.4	–1.1

31. The following table shows the annual consumption and disposable income for a certain country (in $ million).

Year	C	Y_d	Year	C	Y_d
1957	11,378	11,617	1963	20,074	21,512
1958	13,012	13,297	1964	21,439	23,124
1959	15,263	15,790	1965	22,833	24,724
1960	16,873	18,017	1966	24,208	26,175
1961	17,764	19,314	1967	25,307	27,219
1962	18,857	20,198	1968	27,020	28,915

Application of OLS yields the following results :

$$\hat{C}_t = 8,526 + 0.65\ Y_{dt},\ R^2 = 0.953$$

(i) Find the residuals and test for autocorrelation.

(ii) Estimate the value of ρ if autocorrelation is of first order (using any method).

Use your estimates of ρ to transform your original data

$$Y_t^* = (Y_t - \hat{\rho} Y_{t-1}),\ X_t^* = (X_t - \hat{\rho} X_{t-1}).$$

Apply OLS to the transformed data and compare your results with the OLS estimates obtained from the original sample observations.

32. Assuming that ρ = 1, compute the OLS estimate with the appropriately transformed data. Compare your results with those obtained in the previous exercise.

33. The following table shows the annual consumption and disposable income of a country (in $ million)

Years	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
C_t	26.1	29.3	35.6	39.4	42.7	46.3	50.1	54.5	60.1	64.9	69.2	71.1
Y_{dt}	38.3	43.5	53.5	60.8	66.4	71.2	77.2	86.1	94.6	102.4	109.9	11[illegible]

(i) Estimate the savings function $S_t = f(Y_{dt})$.

(ii) Test for heteroscedasticity using Spearman's rank correlation coefficient.

(iii) Estimate from the sample data the form of heteroscedasticity.

(iv) Re-estimate the savings function with an appropriately transformed model.

34. The following table shows time series on three variables Y, X_1, X_2 in arbitrary units.

Y	6.0	6.0	6.5	7.1	7.2	7.6	8.0	9.0	9.0	9.3
X_1	40.1	40.3	47.5	49.2	52.3	58.0	61.3	62.5	64.7	66.8
X_2	5.5	4.7	5.2	6.8	7.3	8.7	10.2	14.1	17.1	21.3

(i) Test for multicollinearity with any appropriate method.

(ii) How does multicollinearity affect the parameter estimates ?

35. The following table shows the value of output (X), the labour input (L) and the capital input (K) of 20 firms :

X :	82	73	58	68	98	83	100	110	120	95
L :	15	39	99	12	42	95	45	36	40	65
K :	90	40	20	60	60	30	60	80	80	40
X :	115	64	140	85	56	150	65	36	57	50
L :	30	60	100	95	75	90	25	80	12	65
K :	80	30	60	40	20	90	30	10	40	20

(i) Obtain estimates of a Cobb-Douglas production function using the observations 1-15.

(ii) Explore the pattern of multicollinearity and the effects on the estimates.

(iii) Test the hypothesis that the estimates are sensitive to sample size, utilising the additional information of observations 16–20.

(iv) Comment on your results.

36. Consider the following linear production function of an economy.

$$X_t = \beta_0 + \beta_1 L_t + \beta_2 K_t + u_t$$

where X_t = quantity of output in period t
L_t = labour input (in man hours)
K_t = capital input (in machine hours)

(i) If it is known that technical progress is taking place, discuss the nature and effects of the mis-specification bias imparted in the $\hat{\beta}$ s.

(ii) Suppose K is omitted from the function. Explore the bias imparted in $\hat{\beta}_1$.

37. Consider the model $Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t$

where X's are orthogonal. Show that

(i) the $\hat{\beta}$'s from multiple regression are identical to the coefficient estimates obtained by a simple regression of Y on the corresponding X.

(ii) the total sum of squared residuals is just the sum of the sum of squared residuals of all the simple regressors.

38. (i) Using the hypothetical observations of the following table, show that multicollinearity makes the coefficient of X_1 unstable and increases its standard error.

n	1	2	3	4	5	6	7	8
Y :	9	5	8	6	8	5	9	6
X_1 :	9	4	8	7	8	4	9	7
X_2 :	7	2	4	3	4	2	7	3

(ii) Is this a general result of multicollinearity ? Intermediate results :

$\Sigma Y = 56$, $\Sigma X_1 = 56$, $\Sigma X_2 = 32$, $\Sigma X_1 Y = 414$, $\Sigma X_1^2 = 420$, $\Sigma X_2^2 = 156$, $\Sigma X_2 Y = 246$, $\Sigma X_1 X_2 = 248$, $\Sigma Y^2 = 412$.

39. The omission of an important explanatory variable from a function imparts a 'mis-specification bias' in the estimates of the coefficients of the included variables. Derive an algebraic expression to show the nature and the determinants of this bias.

40. Consider the model (in deviation form) $y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$ $(i = 1, 2, \ldots, n)$. If X_1 and X_2 are orthogonal (*i.e.*, $r_{X_1 X_2} = 0$), show that the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ obtained from applying multiple regression to the above model are identical with the estimates β_1^* and β_2^* obtained from the simple regressions $y = \beta_1 x_1 + v$ and $y = \beta_2 x_2 + w$ (where v and w are random terms satisfying the usual assumptions)

41. The following table shows the quantity demanded of a certain commodity, its price and consumer's income. The variables are measured in arbitrary units.

Year :	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Demand (Y) :	3.5	4.3	5.0	6.0	7.0	9.0	8.0	10	12	14
Price (X_1) :	16	13	10	7	7	5	4	3	3.5	2
Income (X_2) :	15	20	30	42	50	54	65	72	85	90

(a) Discuss the effects of multicollinearity on the basis of the following results :

(i) $r_{X_1 X_2} = -0.943$

(ii) $Y_t = \beta_0 + \beta_1 X_{1t} + u_t$

$\hat{Y}_t = 12.4879 - 0.6536 X_{1t}$

SE : (0.121), $R^2 = 0.784$

t : (−5.38), $F = 28.97$

(iii) $Y_t = \beta_0 + \beta_2 X_{2t} + u_t$

$\hat{Y}_t = 1.2179 + 0.12738 X_{2t}$

SE : (0.011), $R^2 = 0.942$

t : (11.44), $F = 130.76$

(iv) $Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t$

$\hat{Y}_t = -1.9194 + 0.11841 X_{1t} + 0.16062 X_{2t}$

SE : (0.19) (0.03), $R^2 = 0.950$

$F = 67.04$

(b) Explain the specification bias in model (ii) in which "income" is omitted, and in model (iii) in which 'price' is omitted.

(c) Would you omit X_2 from the model ?

5

Specification Analysis

5.1. Introduction

The early developments in econometrics were concerned with the problems of estimation of an econometric model once the model was specified (in which all the possible relevant explanatory variables were included in the model and irrelevant variables were excluded from the model). The major preoccupation of econometricians was with devising methods of estimation that produced consistent and efficient estimates of the parameters. These are the methods which we have discussed in the earlier chapters.

During recent years, the attention of econometricians has been directed to the problems of :

1. Checking the adequacy of the specification of the model. This is called "diagnostic checking" and "specification testing".
2. Choosing between alternative specifications of the model. This is called "model selection".
3. Devising methods of estimation based on weaker assumptions about the error distributions. This is called "Semi Parametric estimation". However, the last one is beyond the scope of this book. The first two areas are also very vast in scope.

In fact, we have already discussed some tests for diagnostic checking in earlier chapters. For instance :

(i) Tests for parameter stability

(ii) Tests for heteroscedasticity

(iii) Tests for autocorrelation.

However, these tests are all based on the least squares residuals and during recent years some alternative residuals have been suggested. Also, tests for diagnostic checking have been more systematized.

5.2. Diagnostic Tests Based on Least Squares Residuals

Diagnostic tests are tests that are meant to "diagnose" some problems with the models that we are estimating. The least squares residuals play an important role in many diagnostic tests. We have already discussed about the tests relating to parameter stability, heteroscedasticity and autocorrelation.

We are now considering some other tests based on least squares residuals.

One of the assumptions of the Classical Linear Regression Model (CLRM) is that the regression model used in the analysis is "correctly" specified (including all the relevant variables and excluding all irrelevant variables). If the model is not "correctly" specified we encounter the problem of model specification error or model specification bias.

5.3. Model Selection Criteria

Following Hendry and Richard, a model chosen for empirical analysis should satisfy the following criteria :

1. Predictions made from the model must be logically possible.
2. It must make good economic sense. For instance, in testing *Permanent income hypothesis* of Milton Friedman the intercept value in the regression of permanent consumption on permanent income is expected to be zero.
3. The explanatory variables or regressors must be uncorrelated with the error term.
4. The values of the parameters should be stable, otherwise, forecasting will be difficult. In the absence of parameter constancy, predictions will not be reliable.
5. The residuals estimated from the model must be purely random. If it is not so, there will be some specification error in the model.
6. The model should include all the rival models in the sense that it is capable of explaining their results. In short, other models cannot be an improvement over the chosen model. This means that the best fitting model should be chosen.

5.4. Types of Specification Errors

A good econometric model should satisfy the above listed criteria.

Let us consider a model,

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + u_{1i} \text{ .... } (1)$$

where Y = total cost of production and X = output. Equation (1) is the familiar text book example of the cubic total cost function.

But suppose a researcher decides to use the following model :

$$Y_i = \alpha_0 + \alpha_1 X_i + \alpha_2 X_i^2 + u_{2i} \text{ .... } (2)$$

It should be noted that we have changed the notations to distinguish model (2) from model (1). Since equation (1) is assumed to be true, adopting equation (2) would constitute a specification error, the error consists in omiting a relevant variable (X_i^3). Therefore, the error term u_{2i} in equation (2) is in fact

$$u_{2i} = u_{1i} + \beta_3 X_i^3 \text{ .... } (3)$$

Now suppose another researcher uses the following model :

$$Y_i = \lambda_0 + \lambda_1 X_i + \lambda_2 X_i^2 + \lambda_3 X_i^3 + \lambda_4 X_i^4 + u_{3i} \text{ .... } (4)$$

If equation (1) is true, equation (4) also constitutes a specification error, the error here consists in including an unnecessary or irrelevant variable in the sense that the true model assumes λ_4 to be zero. The new error term is in fact,

$$u_{3i} = u_{1i} - \lambda_4 X_i^4 = u_{1i} \text{ .... } (5), \text{ since } \lambda_4 = 0 \text{ in true model (1).}$$

Now assume that another researcher postulates the following model :

$$\log Y_i = \gamma_0 + \gamma_1 X_i + \gamma_2 X_i^2 + \gamma_3 X_i^3 + u_{4i} \text{ .... } (6)$$

In relation to the true model [equation (1)], equation (6) would also constitute a specification bias, the bias here being the use of wrong functional form. In equation (1) Y appears linearly whereas in equation (6) Y appears log linearly.

Finally, we consider a situation where the researcher uses the following model :

$$Y_i^* = \beta_0^* + \beta_1^* X_i^* + \beta_2^* X_i^{2*} + \beta_3^* X_i^{3*} + u_i^* \text{ ....(7)}$$

where $Y_i^* = Y_i + \varepsilon_i$ and $X_i^* = X_i + w_i$, ε_i and w_i being the errors of measurement.

Equation (7) states that instead of using the true Y_i and X_i we have used their proxies, Y_i^* and X_i^*, which may contain errors of measurement. Therefore, in equation (7) we commit the **errors of measurment bias.**

In applied work data are plagued by errors of approximations or errors of incomplete coverage or simply errors of omitting some observations. In social sciences we often depend on secondary data and usually have no way of knowing the types of errors, made by the primary data collecting agency.

Another type of specification error relates to the way the stochastic error u_i (or u_t) enters the regression model. For example, we consider the following bivariate regression model without the intercept term : $Y_i = \beta X_i \mu_i$ .... (8) where the stochastic error term enters multiplicatively with the property that $\log u_i$ satisfies the assumptions of the CLRM against the following model :

$Y_i = \alpha X_i + u_i$ .... (9) where the error term enters additively. Although the variables are the same in the two models, we have denoted the slope coefficient in equation (8) by β and the slope coefficient in equation (9) by α. Now if equation (8) is the 'correct' or 'true' model, would the estimated α provide an unbiased estimate of the true β ? That is, will $E(\hat{\alpha}) = \beta$? If that is not the case, improper stochastic specification of the error term will constitute another source of specification error.

A specification error that is sometimes overlooked is the interaction among the regressors, that is, the multiplicative effect of one or more regressors on the regressand. For example, we consider the following wage function :

$$\log w_i = \beta_0 + \beta_1 \text{ Education} + \beta_2 \text{ Gender} + \beta_3 \text{ (Education)} \times \text{(Gender)} + u_i \text{ ....(10)}$$

In this model, the change in the relative wages with respect to education depends not only on education but also on gender $\left(\dfrac{\partial \log w}{\partial \text{ Education}} = \beta_1 + \beta_3 \cdot \text{Gender}\right)$. Similarly, the change in relative wages with respect to gender depends not only on gender but also on education.

To sum up, in developing an empirical model, one is likely to commit one or more of the following specification errors :

1. Omission of relevant variable (s)
2. Inclusion of unnecessary / irrelevant variable(s).
3. Adoption of the wrong functional form
4. Errors in measurement
5. Incorrect specification of the stochastic error term
6. Assumption that the error term is normally distributed.

The first four types of errors are essentially in the nature of **model specification** errors in the sense that we have in mind a "true" model but somehow we do not estimate the correct model. In the last two cases we may commit **model mis-specification errors** where we do not know what the true model is to begin with.

In the present book we will concentrate mainly on model specification errors.

5.5. Consequences of Model Specification Errors

For the sake of simplicity we will discuss the consequences of model specification errors in terms of three variable regression models and we will discuss the first two types of specification errors, namely,

1. Underfitting a model, that is, omitting relevant variable(s),

and 2. Overfitting a model, that is, including unnecessary/irrelevant variable(s).

5.5.1. Underfitting a Model (Omitting a Relevant Variable)

Suppose the true model is :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \text{ ....(1)}$$

But for some reason we fit the following model :

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + v_i \text{ ....(2)}$$

This means that we have omitted the variable X_2 from model (1).

The consequences of omitting variable X_2 are as follows :

(i) If the omitted variable X_2 is correlated with the included variable X_1, i.e., if $r_{X_1X_2}$ is non zero, then $\hat{\alpha}_0$ and $\hat{\alpha}_1$ are biased as well as inconsistent.

This means that $E(\hat{\alpha}_0) \neq \beta_0$ and $E(\hat{\alpha}_1) \neq \beta_1$ and the bias does not disappear as the sample size gets larger.

(ii) Even if X_1 and X_2 are not correlated, $\hat{\alpha}_0$ is biased, although $\hat{\alpha}_1$ is not biased.

(iii) The disturbance variance σ^2 is incorrectly estimated.

(iv) The conventionally measured variance of $\hat{\alpha}_1 \left(= \dfrac{\sigma^2}{\Sigma x_{1i}^2}\right)$ is a biased estimator of $\hat{\beta}_1$.

(v) The usual confidence interval and hypothesis testing procedures are likely to give misleading conclusions about the statistical significance of the estimated parameters.

(vi) The forecasts based on the incorrect model and the forecast intervals will be unreliable.

Proof : The proofs of some of the above statements are given below :

In the deviation form the three variable population regression model can be written as

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + (u_i - \bar{u}) \text{ ....(1)'}$$

[where $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$

$$\therefore \bar{Y} = \beta_0 + \beta_1 \bar{X}_1 + \beta_2 \bar{X}_2 + \bar{u}$$

or, $$Y_i - \bar{Y} = \beta_1 (X_{1i} - \bar{X}_1) + \beta_2 (X_{2i} - \bar{X}_2) + (u_i - \bar{u})$$

or, $$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + (u_i - \bar{u})]$$

First multiplying by x_{1i} and then by x_{2i}, the usual normal equations become :

$$\Sigma x_{1i} y_i = \beta_1 \Sigma x_{1i}^2 + \beta_2 \Sigma x_{1i} x_{2i} + \Sigma x_{1i} (u_i - \bar{u}) \text{ ....(2)'}$$

and $$\Sigma x_{2i} y_i = \beta_1 \Sigma x_{1i} x_{2i} + \beta_2 \Sigma x_{2i}^2 + \Sigma x_{2i} (u_i - \bar{u}) \text{ ....(3)'}$$

Now dividing equation (2)' by Σx_{1i}^2 on both sides we have,

$$\frac{\Sigma x_{1i}y_i}{\Sigma x_{1i}^2} = \beta_1 + \beta_2 \frac{\Sigma x_{1i}x_{2i}}{\Sigma x_{1i}^2} + \frac{\Sigma x_{1i}(u_i - \bar{u})}{\Sigma x_{1i}^2} \text{ ....(4)'}$$

Since from model (2) we can have

$$\hat{\alpha}_1 = \frac{\Sigma x_{1i}y_i}{\Sigma x_{1i}^2} \text{ [where } y_i = \alpha_1 x_{1i} + (v_i - \bar{v})\text{, in deviation form]}$$

and again $b_{21} = \dfrac{\Sigma x_{1i}x_{2i}}{\Sigma x_{1i}^2}$ = Regression coefficient of X_2 on X_1 i.e., the slope in the regression of excluded variable X_2 on the included variable X_1.

Hence, from equation (4)' we have

$$\hat{\alpha}_1 = \beta_1 + \beta_2 b_{21} + \frac{\Sigma x_{1i}(u_i - \bar{u})}{\Sigma x_{1i}^2} \text{ ....(5)'}$$

Taking the expected value of the equation (5)' we finally obtain

$$E(\hat{\alpha}_1) = \beta_1 + \beta_2 b_{21} \text{ ....(6)'}$$

where (a) b_{21} is a known fixed quantity,

(b) β_1 and β_2 are constants,

and (c) u_i is uncorrelated with X_{1i} (as well as X_{2i}).

so, $E(\hat{\alpha}_1) = \beta_1 + \beta_2 b_{21} \neq \beta_1$

Therefore $\hat{\alpha}_1$ is a biased estimator of β_1 unless β_2 or b_{21} or both are zero.

We rule out β_2 being zero because in that case we do not have specification error to begin with.

If, however, $b_{21} = 0$ i.e. X_1 and X_2 are uncorrelated, then $E(\hat{\alpha}_1) = \beta_1$. But this is also unlikely in most economic data.

Thus, bias is given by,

$$E(\hat{\alpha}_1) - \beta_1 = \beta_2 \cdot b_{21}$$

= (coefficient of the excluded variable) × (regression coefficient in a regression of the excluded variable on the included variable).

Generally, the extent of the bias $[E(\hat{\alpha}_1) - \beta_1 = \beta_2 \cdot b_{21}]$ will depend on the bias term $\beta_2 b_{21}$. If, for instance, β_2 is positive (i.e., X_2 has a positive effect on Y) and b_{21} is positive (i.e., X_1 and X_2 are positively correlated), $\hat{\alpha}_1$, on an average will overestimate the true β_1 (i.e., positive bias). But this result should not be surprising, for X_1 represents not only its direct effect on Y but also its indirect effect (via X_2) on Y.

In short, X_1 gets credit for the influence that is rightly attributable to X_2, the latter being prevented from showing its effect explicitly because it is not 'allowed' to enter the model.

Variances of $\hat{\alpha}_1$ and $\hat{\beta}_1$:

Let us now examine the variances of $\hat{\alpha}_1$ and $\hat{\beta}_1$.

Conventionally, $\text{var}(\hat{\alpha}_1) = \dfrac{\sigma^2}{\Sigma x_{1i}^2}$ ....(7)'

and $\text{var}(\hat{\beta}_1) = \dfrac{\sigma^2}{\Sigma x_{1i}^2\left(1 - r_{X_1X_2}^2\right)} = \dfrac{\sigma^2}{\Sigma x_{1i}^2} \cdot \text{VIF}$ .... (8)'

[Here we write $\sigma_v^2 = \sigma^2$ for the sake of simplicity.]

where VIF (a measure of collinearity) is the variance inflation factor $\left[= \dfrac{1}{\left(1 - r_{X_1X_2}^2\right)}\right]$ [See 4.14.3] and $r_{X_1X_2}$ is the correlation coefficient between X_1 and X_2.

It should be noted that as formulas (7)' and (8)' are not the same, in general, $\text{var}(\hat{\alpha}_1)$ will be different from $\text{var}(\hat{\beta}_1)$. Since $0 < r_{X_1X_2}^2 < 1$, it would imply $\text{var}(\hat{\alpha}_1) < \text{var}(\hat{\beta}_1)$.

Although $\hat{\alpha}_1$ is biased, its variance is smaller than the variance of the unbiased estimator $\hat{\beta}_1$ (of course, we are ruling out the case where $r_{X_1X_2} = 0$, since in practice there is some correlation between regressors). So, there is a trade off involved here.

This is not the end of the story because σ^2 estimated from the model $Y_i = \alpha_0 + \alpha_1 X_{1i} + v_i$ and that estimated from the true model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ are not the same as the residual sum of squares (RSS) of the two models as well as their degrees of freedom (d.f) are different $\left[\dfrac{\Sigma e_i^2}{df} = \dfrac{RSS}{n-K} \text{ and } E\left[\dfrac{\Sigma e_i^2}{n-K}\right] = \hat{\sigma}_u^2\right]$ where K = number of parameters, n = sample size]. Now if we add variables to the model, the RSS generally decreases but the *degrees of freedom* $(n-K)$ also decrease because more parameters are estimated. The net outcome depends on whether the RSS decreases sufficiently to offset the loss of degrees of freedom due to the addition of regressors. It is quite possible that if a regressor has a strong impact on the regressand, inclusion of variables in the model will not only reduce the bias but will also increase the precision (i.e., reduce the standard errors) of the estimates. On the other hand, if the relevant variables have only a marginal impact on the regressand and if they are highly correlated (i.e., $\text{VIF} = \dfrac{1}{1 - r_{X_1X_2}^2}$ is larger), we may reduce the bias in the coefficients of the variables already included in the model, but increase their standard errors (i.e., make them less efficient).

Indeed, the trade-off in this situation between bias and precision can be substantial. The trade off will depend on the relative importance of the various regressors.

Lastly, we consider the special case when $r_{X_1X_2} = 0$. This will result in $b_{21} = 0$. In this case $E(\hat{\alpha}_1) = \beta_1 + \beta_2 \cdot b_{21} = \beta_1$ $\therefore E(\hat{\alpha}_1) = \beta_1$. Here $\hat{\alpha}_1$ is an unbiased estimator of β_1. If $r_{X_1X_2} = 0$, $\text{var}(\hat{\alpha}_1) = \text{var}(\hat{\beta}_1) = \dfrac{\sigma^2}{\Sigma x_{1i}^2}$. This shows that when $r_{X_1X_2} = 0$, the variances of $\hat{\alpha}_1$ and $\hat{\beta}_1$ are the same.

Is there no harm in dropping the variable X_2 from the model even though it may be relevant theoretically ?

The answer generally is no as var $(\hat{\alpha}_1)$ estimated from equation $\text{var}(\hat{\alpha}_1)=\frac{\sigma^2}{\Sigma x_{1i}^2}$ is still biased and therefore our hypothesis testing procedures are likely to remain suspect. Besides, in most economic research X_1 and X_2 will be correlated. The point is clear : Once a model is formulated on the basis of the relevant theory, one is ill advised to drop a variable from such a model.

Example 5.1. *On the basis of cross-sectional data for 64 countries on child mortality and a few other variables the following linear regression model has been fitted :*

$$CM_i=\beta_0+\beta_1 PGNP_i+\beta_2 FLR_i+u_i$$

where CM = child mortality, the number of deaths of children under age 5 in a year per 1000 live births, FLR = Female literacy rate, percent,

PGNP = per capita GNP in 1980

$i=1, 2, ..., 64$ as there are 64 countries.

The estimated regression results were obtained as follows :

$$\hat{CM}_i=263.6416-0.0056\,PGNP_i-2.2316\,FLR_i$$

$$SE:\ (11.5932)\quad (0.0019)\qquad\qquad (0.2019)\qquad R^2=0.7077,\ \bar{R}^2=0.6981$$

Let us now interpret the regression coefficients : –0.0056 is the partial regression coefficient of PGNP and tells us that with the influence of FLR held constant, as PGNP increases, say, by a dollar on an average, child mortality goes down by 0.0056 units. More specifically we can say that if the percapita GNP goes up by a thousand dollars, on an average, the number of deaths of children under age 5 goes down by about 5.6 per thousand live births. The coefficient –2.2316 tells us that holding the influence of PGNP constant, on an average the number of deaths of children under age 5 goes down by about 2.23 per thousand live births as the female literacy rate increases by one percentage point. The intercept value of about 263 means that if the PGNP and FLR were fixed at zero, the mean child mortality rate would be about 263 deaths per thousand live births.

The R^2 value of about 0.71 means that about 71 percent of the variation in child mortality is explained by PGNP and FLR. All the regression results make sense. Further more, all the regression paramaters are statistically significant individually.

Assuming the above stated model as the correct or true model we drop the FLR variable from the model and examine its consequence. If we drop FLR from the true model and regress the given data set we obtain the coefficient of PGNP (in incorrect model) –0.0114 but in the correct/true model it was –0.0056.

In absolute terms, now PGNP has a greater impact on CM as compared with the true model. But if we regress FLR on PGNP (Regression of the excluded variable on the included variable), the slope coefficient in the regression (b_{21}) becomes 0.00256. The regression results are :

$$\hat{FLR}_i=47.5971+0.00256\,PGNP_i$$

$$SE:\quad (3.5553)\quad (0.0011),\qquad\qquad r^2=0.0721$$

This suggests that as *PGNP* increases by a unit, on an average, *FLR* goes up by 0.00256 units. But if *FLR* goes up by these units, its effect on *CM* will be $\hat{\beta}\text{-}b_{21}=$ $-2.2316\times 0.00256=-0.00543$. Therefore, we finally have,

$\hat{\beta}_1+\hat{\beta}_2\cdot b_{21}=-0.0056+(-2.2316)\times(0.00256)\approx-0.0114$ which is about the value of the *PGNP* coefficient obtained in the true model. This example clearly shows that the true impact of *PGNP* on *CM* is much less (–0.0056) than that suggested by the incorrect model, namely (–0.0114).

5.5.2. Inclusion of an Irrelevant Variable (Overfitting a Model)

To begin with we consider a model

$$Y_i=\beta_0+\beta_1 X_{1i}+u_i \text{ .... (1)}$$

which is the correct/true model but we fit the following (incorrect) model,

$$Y_i=\alpha_0+\alpha_1 X_{1i}+\alpha_2 X_{2i}+v_i \text{ .... (2)}$$

in which we have included an unnecessary/irrelevant variable X_2 and hence we commit the specification error.

The consequences of this specification error are as follows :

(*i*) The *OLS* estimators of the parameters of the 'incorrect' model are all unbiased and consistent ; i.e. $E(\hat{\alpha}_0)=\beta_0$, $E(\hat{\alpha}_1)=\beta_1$ and $E(\hat{\alpha}_2)=\beta_2=0$.

(*ii*) The error variance σ^2 is correctly estimated.

(*iii*) The usual confidence interval and hypothesis testing procedures remain valid.

(*iv*) However, the estimated α's will be generally inefficient, that is, their variances will be generally larger than those of the β's of the true model.

Proof : The proofs of some of the above statements are given below :

For the true model, $Y_i=\beta_0+\beta_1 X_{1i}+u_i$ .... (1)

Usually we have $\hat{\beta}_1=\frac{\Sigma x_{1i}y_i}{\Sigma x_{1i}^2}$ where $x_{1i}=X_{1i}-\bar{X}_1$, $y_i=Y_i-\bar{Y}$ and also we know that $\hat{\beta}_1$ is an unbiased estimator of β_1. For the overfitting model (after the inclusion of irrelevant variable X_2),

$Y_i=\alpha_0+\alpha_1 X_{1i}+\alpha_2 X_{2i}+v_i$ ________ (2) we usually obtain (by OLS),

$$\hat{\alpha}_1=\frac{\Sigma x_{1i}y_i\cdot\Sigma x_{2i}^2-\Sigma x_{2i}y_i\cdot\Sigma x_{1i}x_{2i}}{\Sigma x_{1i}^2\cdot\Sigma x_{2i}^2-(\Sigma x_{1i}x_{2i})^2}$$

and $$\hat{\alpha}_2=\frac{\Sigma x_{2i}y_i\cdot\Sigma x_{1i}^2-\Sigma x_{1i}y_i\cdot\Sigma x_{1i}x_{2i}}{\Sigma x_{1i}^2\cdot\Sigma x_{2i}^2-(\Sigma x_{1i}x_{2i})^2}$$

where $x_{1i}=X_{1i}-\bar{X}_1$, $x_{2i}=X_{2i}-\bar{X}_2$, and $y_i=Y_i-\bar{Y}$.

Now from the true model, $Y_i=\beta_0+\beta_1 X_{1i}+u_i$ .... (1)

we have $\bar{Y}=\beta_0+\beta_1\bar{X}_1+\bar{u}$

$$\text{or, } Y_i-\bar{Y}=\beta_1(X_{1i}-\bar{X}_1)+(u_i-\bar{u})$$

$$\text{or, } y_i=\beta_1 x_{1i}+(u_i-\bar{u})$$

We now put this in the expression of $\hat{\alpha}_1$ and get

$$\hat{\alpha}_1 = \frac{\Sigma x_{1i}[\beta_1 x_{1i} + (u_i - \bar{u})]\,\Sigma x_{2i}^2 - \Sigma x_{2i}[\beta_1 x_{1i} + (u_i - \bar{u})]\Sigma x_{1i} x_{2i}}{\Sigma x_{1i}^2\, \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2}$$

$$\therefore E(\hat{\alpha}_1) = \frac{\beta_1[\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2]}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2}, \text{ since } E(u_i - \bar{u}) = 0$$

or, $E(\hat{\alpha}_1) = \beta_1$. This proves that $\hat{\alpha}_1$ is an unbiased estimator of β_1.

Similarly, from the expression of $\hat{\alpha}_2$ [after letting $y_i = \beta_1 x_{1i} + (u_i - \bar{u})$] we have

$$\hat{\alpha}_2 = \frac{\Sigma x_{2i}[\beta_1 x_{1i} + (u_i - \bar{u})]\Sigma x_{1i}^2 - \Sigma x_{1i}[\beta_1 x_{1i} + (u_i - \bar{u})]\Sigma x_{1i} x_{2i}}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2}$$

$$\therefore E(\hat{\alpha}_2) = \frac{\beta_1[\Sigma x_{1i} x_{2i} \cdot \Sigma x_{1i}^2 - \Sigma x_{1i} x_{2i} \cdot \Sigma x_{1i}^2]}{\Sigma x_{1i}^2 \cdot \Sigma x_{2i}^2 - (\Sigma x_{1i} x_{2i})^2}, \text{ as } E(u_i - \bar{u}) = 0$$

$$= 0$$

$$\therefore E(\hat{\alpha}_2) = 0 = \beta_2$$

which is its value in the true model where X_2 is absent.

From the true model $Y_i = \beta_0 + \beta_1 X_{1i} + u_i$

we have (by usual OLS formula), $\text{var}(\hat{\beta}_1) = \dfrac{\sigma^2}{\Sigma x_{1i}^2}$ and from the incorrect model (after the inclusion of irrelevant variable X_2), $Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + v_i$

we have $\text{var}(\hat{\alpha}_1) = \dfrac{\sigma^2}{\Sigma x_{1i}^2\left(1 - r_{X_1 X_2}^2\right)}$.

Therefore, $\dfrac{\text{var}(\hat{\alpha}_1)}{\text{var}(\hat{\beta}_1)} = \dfrac{1}{1 - r_{X_1 X_2}^2}$.

Since $0 \le r_{X_1 X_2}^2 \le 1$, it follows that $\text{var}(\hat{\alpha}_1) \ge \text{var}(\hat{\beta}_1)$, *i.e.*, the variance of $\hat{\alpha}_1$ is *generally greater than the variance of* $\hat{\beta}_1$, *even though, on an average* $\hat{\alpha}_1 = \beta_1$ [*i.e.* $E(\hat{\alpha}_1) = \beta_1$]. The implication of this result is that the inclusion of the unnecessary / irrelevant variable X_2 in the model makes the variance of $\hat{\alpha}_1$ greater than necessary, thereby making $\hat{\alpha}_1$ less precise. This is also true for $\hat{\alpha}_0$.

Note : *From the above discussion (5.5.1 & 5.5.2) we observe an asymmetry between two types of biases. If we exclude a relevant variable (5.5.1), the coefficients of the variables retained in the model are generally biased and inconsistent, the error variance is incorrectly estimated, and usual hypothesis testing procedures become invalid. On the other hand, including an irrelevant variable in a model (Sec 5.5.2) still gives us unbiased and consistent estimates of the coefficients in the true model, the error* variance is correctly estimated, and the conventional hypothesis testing methods are still valid ; the only penalty we pay for the inclusion of the superfluous variable is that the estimated variance of the coefficients are larger, and hence less precise. We may thus conclude that it is better to include irrelevant variables than to omit the relevant variables. But the addition of irrelevant variables will lead to a loss in the efficiency of the estimators and may also lead to the problem of multicollinearity.

Therefore, in general, the best approach is to include only explanatory variables that on theoretical grounds, directly influence the dependent variable and that are not accounted for by other included variables.

5.6. Tests of Specification Errors

We have already analysed the consequences of specification errors. Very often specification errors arise inadvertently, mainly from our inability to formulate the model as precisely as possible because the underlying theory is weak or because we do not have the right kind of data to test the model.

Whatever may be the reason or source of specification errors, we like to detect them. Once it is found that specification errors have been made, the remedies often suggest themselves.

For example, if it is found that a variable is inappropriately omitted from a model, the obvious remedy is to include that variable in the analysis, assuming, of course,the data on that variable are available. Now we like to discuss some tests that we can use to detect specification errors.

5.6.1. Detecting the Presence of Unnecessary/ Irrelevant Variables (Overfitting a Model)

Suppose we develop a general linear regression model with K explanatory variables to explain a phenomenon : The regression equation takes the form :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_K X_{Ki} + u_i \quad \ldots\ldots\ldots (1)$$

However, we are not totally sure that, say, the variable X_K really belongs to the model. One simple way to find this out is to test the significance of the estimated β_K with the usual 't' test : $t = \dfrac{\hat{\beta}_K}{SE(\hat{\beta}_K)}$. But suppose that we are not sure whether, say, X_2 and X_3 legitimately belong to the model. This can be easily ascertained by the F test as discussed in Chapter-3 [where we usually test the hypothesis, say, $H_0 : \beta_1 = \beta_2 = \ldots = \beta_K = 0$ against the alternative, $H_1 : \beta_i$'s are not equal to zero].

Thus, detecting the presence of an irrelevant variable is not a difficult task.

It is, however, very important to remember that in carrying out these tests of significance we have a specific model in mind.

We accept the model as a 'true' model. Given that model, we can find out whether one or more regressors are really relevant by the usual 't' (for testing the significance of the parameters individually) and F (jointly) tests. But it may be carefully noted that we should not use the 't' and 'F' tests to build a model '*iteratively*', that is, we should not say that initially Y is related to X_1 only because $\hat{\beta}_1$ is statistically significant and then expand the model to include X_2 and decide to keep that variable in the model if $\hat{\beta}_2$ turns out to be statistically significant, and so on.

This strategy of building a model is called the **bottom up approach** (starting with a smaller model and expanding it as one goes along) or by the term, **data mining** (other names are **regression fishing, data grubbing, data snooping** and **number crunching**).

The primary objective of data mining is to develop the 'best' model after several diagnostic tests so that the model finally chosen is a "good" model in the sense that all the estimated coefficients have the "right" signs, they are statistically significant on the basis of '*t*' and '*F*' tests, the R^2 value is reasonably high and the Durbin-Watson '*d*' has acceptable value (around 2), etc.

5.6.2. Tests for Omitted Variables and Incorrect Functional Form

In practice we are never sure that the model adopted for empirical testing is "the truth, the whole truth and nothing but the truth". On the basis of theory or introspection and prior empirical work, we develop a model that we believe captures the essence of the subject under study. We then subject the model to empirical testing. After obtaining the results, we begin the post-mortem, keeping in mind the criteria of a good model. In order to determine the adequacy of the model, we look at some broad features of the derived results, such as the $\bar{R}^2$ value, the estimated '*t*' ratios, the signs of the estimated coefficients in relation to their prior expectations, the Durbin-Watson statistic, and the like. If these diagnostics are reasonably good, we proclaim that the chosen model is a fair representation of reality. On the other extreme if the results do not look encouraging because the $\bar{R}^2$ value may be too low or because very few coefficients are statistically significant or have the incorrect signs or because the Durbin-Watson '*d*' is too low, then we doubt about the adequacy of the model and search for remedies. May be we have omitted an important variable or have used the wrong functional form, and so on. In order to determine whether model adequacy is on account of one or more of these problems, we can use some of the following methods.

(i) Examination of Residuals

Examination of residuals is a good visual diagnostic test to detect autocorrelation or heteroscedasticity (as discussed in Chapter 4). But these residuals can also be examined, especially in cross-sectional data for model specification errors, such as omission of an important varisble or incorrect functional form. If, in fact, there are such errors, a plot of the residuals will exhibit distinct patterns. To explain this we consider a cubic total cost function. Assume that the 'true' total cost function is of the form :

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + u_{1i} \text{ .... (1)}$$

but a researcher fits the following quadratic cost function :

$$Y_i = \alpha_0 + \alpha_1 X_i + \alpha_2 X_i^2 + u_{2i} \text{ .... (2)}$$

and another researcher fits the following linear cost function :

$$Y_i = \lambda_0 + \lambda_1 X_i + u_{3i} \text{ .... (3)}$$

where Y = total cost, X = output.

If (1) is the 'true' cost function then both the researchers [(2) & (3)] have made specification errors.

Let us consider an example.

Example 5.2. We have the following set of data on output (X) and total cost (Y) :

Output (X) :	1	2	3	4	5	6	7	8	9	10
Total cost (Y) in $:	193	226	240	244	257	260	274	297	350	420

Fit all the three cost functions (cubic, quadratic and linear) on the basis of the above set of data and also estimate the residuals of each model and plot them in a single diagram.

Solution : We have fitted the cost functions,

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + u_{1i} \text{ .... (1)}$$

$$Y_i = \alpha_0 + \alpha_1 X_i + \beta_2 X_i^2 + u_{2i} \text{ .... (2)}$$

$$Y_i = \lambda_0 + \lambda_1 X_i + u_{3i} \text{ .... (3)}$$

and obtained the following regression results (detailed calculations are not shown here) :

$$* \quad \hat{Y}_i = 166.467 + 19.933\, X_i \qquad R^2 = 0.8409$$

$$SE : (19.021) \quad (3.066) \qquad \bar{R}^2 = 0.8210$$

$$t : (8.752) \quad (6.502) \qquad d = 0.716$$

Linear cost function, i.e., model (3)

where D.W statistic,

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

$$** \quad \hat{Y}_i = 222.383 - 8.0250\, X_i + 2.542\, X_i^2 \qquad R^2 = 0.9284$$

Quadratic cost function

$$SE : (23.488) \quad (9.809) \quad (0.869) \qquad \bar{R}^2 = 0.9079$$

i.e., Model (2).

$$t : (9.468) \quad (-0.818) \quad (2.925) \qquad d = 1.038$$

$$*** \quad \hat{Y}_i = 141.767 + 63.478\, X_i - 12.962\, X_i^2 + 0.939\, X_i^3 \qquad R^2 = 0.9983$$

.... Cubic cost function, i.e. Model (1)

$$SE : (6.375) \quad (4.778) \quad (0.9856) \quad (0.0592) \qquad \bar{R}^2 = 0.9975$$

$$t : (22.238) \quad (13.285) \quad (-13.151) \quad (15.861) \qquad d = 2.70$$

The estimated residuals of all the three cost functions are shown in the following table (Table 5.1) :

Table 5.1.

Estimated residuals from the Linear, Quadratic and Cubic Total cost functions

Observation (output)	e_i (Linear Model*)	e_i (Quadratic Model**)	e_i (Cubic Model ***)
1	6.600	–23.900	–0.222
2	19.667	9.500	1.607
3	13.733	18.817	–0.915
4	–2.200	13.050	–4.426
5	–9.133	11.200	4.435
6	–26.067	–5.733	1.032
7	–32.000	–16.750	0.726
8	–28.933	–23.850	–4.119
9	4.133	–6.033	1.859
10	34.200	23.700	0.022

In each case $e_i = Y_i - \hat{Y}_i$, $i = 1, 2, \ldots, 10$

In the following diagram (Fig. 5.1) we have depicted the residuals of the three models, namely, linear, quadratic and cubic cost functions.

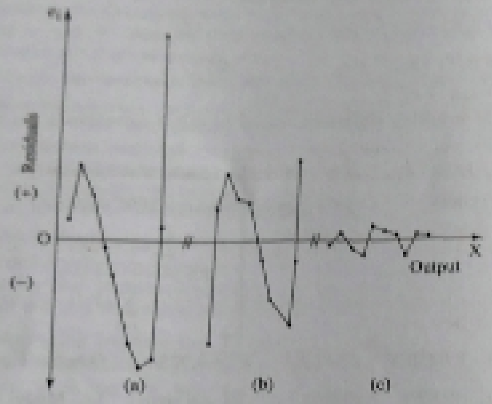


Fig. 5.1. Residuals e_i from (a) Linear, (b) Quadratic, (c) Cubic total cost functions

From the diagram (Fig. 5.1) we see that as we move from left to right, i.e., as we approach the truth, not only are the residuals smaller (in absolute value) but also they do not exhibit the pronounced cyclical swings associated with the misfitted models.

The utility of examining the residual plot is thus clear. If there are specification errors, the residuals will exhibit noticeable patterns.

(ii) The Durbin-Watson '*d*' Statistic

The Durbin-Watson Statistic '*d*' is usually computed by the formula :

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \quad \text{[See Chapter 4]}$$

where $e_t = Y_t - \hat{Y}_t$ =observed value of Y – estimated value of Y.

If we examine the routinely calculated Durbin-Watson Statistic '*d*' of different cost models (linear, quadratic and cubic) we see that for the linear cost function the estimated '*d*' is 0.716, suggesting that there is positive 'correlation' in the estimated residuals : for $n = 10$ and $k' = 1$ (number of explanatory variables excluding the constant term), 5% critical '*d*' values are $d_L = 0.870$ and $d_U = 1.320$. Similarly, the computed '*d*' value for the quadratic cost function is 1.038 whereas the 5% critical '*d*' values are (for $n = 10$, and $k' = 2$), $d_L = 0.69$ and $d_U = 1.64$, indicating indecision (inconclusive) i.e. when $d_L < d < d_U$, the test is inconclusive. Again, the computed '*d*' value for the cubic cost function is $d = 2.70$ where as the 5% critical values are (for $k' = 3$), $d_L = 0.52$, $d_U = 2.01$. Here we see that $d_U < d < (4 - d_L)$ i.e. $2.01 < d = 2.70 < 4 - 0.52 = 3.48$.

So, we accept the hypothesis of no autocorrelation. Here, the estimated '*d*' value does not indicate any positive 'correlation' in the residuals. The observed positive 'correlation' in the residuals in a fitted model is not a measure of (first order) serial correlation (autocorrelation) but of (model) specification error (s).

In particular if we exclude (omit) X_i^3 from the cubic cost function,

$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + u_{1i}$, then error term in the misspecified model

$(Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_{2i})$ would be $u_{2i} = u_{1i} + \beta_3 X_i^3$ and it will exhibit a systematic pattern (e.g., positive autocorrelation) if X_i^3 in fact affects Y significantly.

To use the Durbin-Watson test for detecting model specification error(s), we proceed as follows :

(*i*) From the assumed model, obtain the ordinary least squares (OLS) residuals $[e_i = Y_i - \hat{Y}_i$ or $e_t = Y_t - \hat{Y}_t]$

(*ii*) Ensure that the model is mis-specified as it excludes a relevant explanatory variable from the model.

(*iii*) Compute the '*d*' statistic from the residuals thus ordered by the usual '*d*' formula,

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

. Here the subscript '*t*' is the index of observation here and does not necessarily mean that the data are time series. It can also be applied for cross-sectional data.

(*iv*) From the Durbin-Watson tables, if the estimated '*d*' value is significant, then one can accept the hypothesis of model mis-specification.

If the model mis-specification is detected, the remedial measures will naturally suggest themselves.

In our cost function example the remedies are clear : Indroduce the quadratic term in the linear cost function and the cubic term in the quadratic cost function. In short, we finally run the cubic cost model.

(iii) Ramsey's RESET Test

Ramsey has proposed a general test of specification error called RESET (Regression Specification Error Test). Here we are explaining the simplest version of the test. Here also we are explaining the test with our cost output example and assume that the cost function is linear and is of the form :

$$Y_i = \lambda_0 + \lambda_1 X_i + u_{3i} \text{ .... (1)}$$

where Y = total cost and X = output

Now if we plot the estimated residuals $e_i (e_i = Y_i - \hat{Y}_i)$ obtained from this regression against $\hat{Y}_i$, we get the picture shown in Figure 5.2. Although Σe_i and $\Sigma e_i \hat{Y}_i$ are necessarily zero, the residuals in this figure (Fig. 5.2) show a pattern in which their mean changes systematically with $\hat{Y}_i$.

Table 5.2

Output (X_i)	Total cost (Y_i) in $	Estimated cost function $\hat{Y}_i = 166.467 + 19.933\ X_i$	$e_i = Y_i - \hat{Y}_i$
1	193	186.400	6.600
2	226	206.333	19.667
3	240	226.267	13.733
4	244	246.200	-2.200
5	257	266.133	-9.133
6	260	286.067	-26.067
7	274	306.000	-32.000
8	297	325.933	-28.933
9	350	345.867	4.133
10	420	365.800	54.200

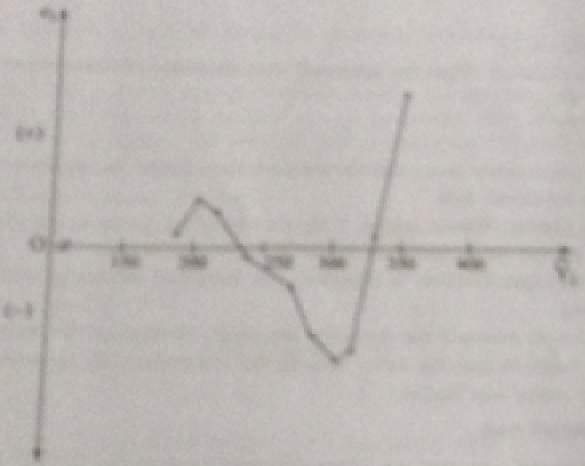


Fig. 5.2. Residuals e_i and estimated Y from the linear cost function $Y_i = \lambda_0 + \lambda_1 X_i + u_i$ (where $\hat{Y}_i = 166.467 + 19.933\ X_i$)

In Figure 5.2 we have plotted e_i (residuals) along the vertical axis (north vertical for positive values and south vertical for negative values) and $\hat{Y}_i$ along horizontal axis respectively.

This would suggest that if we introduce $\hat{Y}_i$ in some form as regressor(s) in the linear cost function, $Y_i = \lambda_0 + \lambda_1 X_i + u_{1i}$, it would increase R^2 (as R^2 increases generally with an increase in the number of regressors) and if the increase in R^2 is statistically significant on the basis of F test, it would suggest that the linear cost function, $Y_i = \lambda_0 + \lambda_1 X_i + u_{1i}$ was mis-specified. This is essentially the idea behind RESET.

The steps involved in RESET are as follows :

(i) From the chosen model, say, $Y_i = \lambda_0 + \lambda_1 X_i + u_{1i}$ we obtain the estimated Y_i that is, $\hat{Y}_i$ .... (1)

(ii) We re-run the regression equation, $Y_i = \lambda_0 + \lambda_1 X_i + u_{1i}$ (linear cost function) introducing $\hat{Y}_i$ in some form as an additional regressor(s). From figure 5.2 we see that there is curvilinear relationship between e_i and $\hat{Y}_i$, suggesting that we can introduce $\hat{Y}_i^2$ and $\hat{Y}_i^3$ as additional regressors. Thus we run the regression equation,

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 \hat{Y}_i^2 + \beta_3 \hat{Y}_i^3 + u_{2i} \text{ .... } (2)$$

(iii) Let the R^2 obtained from equation (2) be denoted by R^2_{new} and that obtained from equation (1) be R^2_{old}.

Then we can use the F test, namely,

$$F = \frac{(R^2_{new} - R^2_{old}) / \text{number of new regressors}}{(1 - R^2_{new}) / (n - \text{number of parameters in the new model})} \text{ .... } (3)$$

to find out if the increase in R^2 from model (2) is statistically significant.

(iv) If the computed F value is significant, say, at the 5% level, we can accept the hypothesis that the linear model (1) is mis-specified.

From our cost output example, the estimated linear cost function (1) was :

$$\hat{Y}_i = 166.467 + 19.933\ X_i$$
$$\text{SE : } (19.021) \quad (3.066), \qquad R^2 = 0.8409$$

Suppose from the new model (2) we run the regression on the basis of data of Table 5.2 (i.e. on the basis of the values of Y_i, X_i and $\hat{Y}_i$) and obtain :

$$\hat{Y}_i = 2140.7223 + 476.6557\ X_i - 0.09187\ \hat{Y}_i^2 + 0.000119\ \hat{Y}_i^3$$
$$\text{SE : } (132.0044) \quad (33.3951) \quad (0.00620) \quad (0.0000074) \qquad R^2 = 0.9983.$$

It should be noted that $\hat{Y}_i^2$ and $\hat{Y}_i^3$ in this equation are obtained from the linear fitted model, $\hat{Y}_i = 166.467 + 19.933\ X_i$

$$\text{SE : } (19.021) \quad (3.066), \qquad R^2 = 0.8409$$

Now applying F test we find,

$$F = \frac{(0.9983 - 0.8409) / 2}{(1 - 0.9983) / (10 - 4)} = 284.4035$$

where $R^2_{new} = 0.9983$, $R^2_{old} = 0.8409$.

$\therefore n = 10$, number of new regressors in the new model = 2, number of parameters in the new model = 4.

Here we see that F value is highly significant, (from table value $F_{0.05\,2,4} = 19.25$) indicating that the model,

$$\begin{array}{llll} \hat{Y}_i = & 166.467 & + \; 19.933\,X_i, & R^2 = 0.8409 \\ \text{SE}: & (19.021) & \quad (3.066) & \end{array}$$

is mis-specified.

Of course, we have reached the same conclusion on the basis of the visual examination of the residuals as well as the Durbin-Watson 'd' value.

However, one advantage of RESET is that it is easy to apply because it does not require one to specify what the alternative model is. But this is also its disadvantage because knowing that a model is mis-specified does not help us necessarily in choosing a better alternative.

(iv) Lagrange Multiplier (LM) Test for Adding Variables

This test is an alternative to Ramsey's RESET test for detecting specification error(s) in a model.

In order to explain this test procedure here also we are using the examples of cost functions. Let $Y_i = \lambda_0 + \lambda_1 X_i + u_{3i}$ .... (1) be a linear cost function, and

$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + u_{1i}$ .... (2) be a cubic cost function where

Y = Total cost and X = output.

If we now compare the linear cost function with the cubic cost function, then the linear cost function will be a restricted version of cubic cost function. The restricted regression $Y_i = \lambda_0 + \lambda_1 X_i + u_{3i}$ assumes that the coefficients of the squared and cubed output terms i.e., β_2 and β_3 are equal to zero.

To test this, the *LM* test proceeds as follows :

(*i*) We estimate the restricted regression $Y_i = \lambda_0 + \lambda_1 X_i + u_{3i}$ by OLS method and obtain the residuals, $e_i = Y_i - \hat{Y}_i$.

(*ii*) If, in fact, the unrestricted regression $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + u_{1i}$ is the true regression, the residuals obtained from restricted regression, $Y_i = \lambda_0 + \lambda_1 X_i + u_{3i}$ should be related to the squared and cubed output terms, i.e., X_i^2 and X_i^3.

(*iii*) This suggests that we regress the e_i obtained in step-I on all the regressors (including those in the restricted regression) which in the present case means

$e_i = \alpha_0 + \alpha_1 X_i + \alpha_2 X_i^2 + \alpha_3 X_i^3 + v_i$ .... (3)

where v is an error term with usual properties.

(*iv*) For large sample size, Engle has shown that n (the sample size) times R^2 estimated from the regression equation (3) follows a chi-square distribution with degrees of freedom (d.f) equal to the number of restrictions imposed by the restricted regression, two in the present example since the terms X_i^2 and X_i^3 are dropped from the model.

Symbolically we write $nR^2 \sim \chi_2^{\ 2}$ .... (4) (no. of restrictions = 2)

(v) If the chi-square value obtained from (4) exceeds the critical value i.e. chi-square value at the chosen level of significance, we *reject the restricted* regression and accept otherwise.

For our example, the regression results of linear *cost function are as follows* :

$$\hat{Y}_i = 166.467 + 19.33X_i \quad \text{..................} (5)$$
$$SE : (19.021) \quad (3.066), \qquad R^2 = 0.8409$$

where Y = total cost, X = output.

When residuals (e_i) from equation (5) are regressed on X_i, X_i^2 and X_i^3 we obtain the following results (where the values of X_i and e_i are taken from Table 5.2) :

$$\hat{e}_i = -24.70 + 43.44X_i - 12.9615X_i^2 + 0.9396X_i^3 \qquad R^2 = 0.9896 \text{ ....} (6)$$
$$SE : (6.375) \quad (4.779) \qquad (0.986) \qquad (0.059)$$

Although our sample size $n = 10$ is by no means large, just *to illustrate the LM* mechanism, we obtain $nR^2 - \chi_2^2 = 10 \times 0.9896 = 9.896$.

From the table value we see that $\chi^2_{0.01;2} = 9.210$ (critical χ^2 value with 2 *degrees of* freedom at 1% level of significance). Thus we see that the *observed value*, $nR^2 - \chi_2^2 = 9.896 > \chi^2_{0.01;2} = 9.210$. So, we reject the restricted regression (*i.e.*, the linear cost function) on the basis of the sample data. We reached a similar conclusion on the basis of Ramsey's RESET test.

From the above discussion on specification errors it is thus clear that we have implicitly assumed that both the dependent variable Y and the explanatory variables, X's are measured without any errors. But in reality the dependent variable Y as well as the explanatory variables i.e., X's may be involved with errors (errors in variables). The detection, consequences and remedies of such errors are beyond the scope of this book. Interested studients are advised to follow standard text books on econometrics for this purpose.

EXERCISE

1. What do you mean by specification problem ? What is meant by diagnostic checking of a model ? Explain it with the help of an example.
2. What do you mean by model selection ? What are the criteria of selecting a good model ?
3. What do you mean by specification errors ? Give a brief outline on different types of specification errors.
4. What do you mean by model specification errors ? What are the consequences of model specification errors ?
5. What do you mean by model specification errors ? Examine the consequences of model specification error(s) when a relevant variable is omitted from a regression model.
6. What do you mean by model mis-specification ? Examine the consequences of model specification errors when an irrelevant varianle is included in a regression model.
7. What do you mean by inclusion of an irrelevant variable in a regression model ? How can you detect the presence of unnecessary / irrelevant variable(s) in a regression model ?
8. What do you mean by exclusion of a relevant variable from a regression model ? What are the different types of tests used for detecting exclusion of relevant variable(s) from a model or incorrect form of the selected model ?

9. Briefly explain (with suitable example) the following methods for determining the adequacy of a model :

(a) Estimation of Residuals

(b) The Durbin-Watson 'd' statistic

(c) Ramsey's RESET test

(d) Lagrange Multiplier (LM) test

10. Suppose the true model is $Y_i = \beta_1 X_i + u_i$ ...... (1) but instead of fitting this regression through the origin you routinely fit the usual intercept-present model :

$Y_i = \alpha_0 + \alpha_1 X_i + v_i$ ...... (2).

Assess the consequence of this specification error.

11. Continue with exercise 10 but assume that the model $Y_i = \alpha_0 + \alpha_1 X_i + v_i$ is the true model. Discuss the consequences of fitting the mis-specified model $Y_i = \beta_1 X_i + u_i$.

12. Suppose that the 'true' model is $Y_i = \beta_0 + \beta_1 X_{1i} + u_i$ ...... (1) but we add an 'irrelevant' variable X_2 to the model (irrelevant in the sense that the true β_2 coefficient attached to the variable X_2 is zero) and estimate $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + v_i$ ...... (2)

(i) Would the R^2 and the adjusted R^2 for model (2) be larger than that for model (1)?

(ii) Are the estimates of β_0 and β_1 obtained from model (2) unbiased ?

(iii) Does the inclusion of the 'irrelevant' variable X_2 affect the variances of $\hat{\beta}_0$ and $\hat{\beta}_1$?

13. Consider the following 'true' Cobb-Douglas production function ;

$\log Y_i = \alpha_0 + \alpha_1 \log L_{1i} + \alpha_2 \log L_{2i} + \alpha_3 \log k_i + u_i$

where Y = output
L_1 = production labour
L_2 = non production labour
k = capital

But suppose the regression actually used in empirical investigation is

$\log Y_i = \beta_0 + \beta_1 \log L_{1i} + \beta_2 \log k_i + u_i$

On the assumption that you have cross-section data on the relevant variables,

(i) will $E(\hat{\beta}_1) = \alpha_1$ and $E(\hat{\beta}_2) = \alpha_3$?

(ii) Will the answer in (i) hold if it is known that L_2 is an irrelevant input in the production function ?

14. Suppose the true model is $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ ..... (1)

but for some reason we fit the following model $Y_i = \alpha_0 + \alpha_1 X_{1i} + v_i$ .... (2)

(i) Find $\text{var}(\hat{\alpha}_1)$ and $\text{var}(\hat{\beta}_1)$.

(ii) Show that $\text{var}(\hat{\alpha}_1) \le \text{var}(\hat{\beta}_1)$

15. Suppose the true model is $Y_i = \beta_0 + \beta_1 X_{1i} + u_i$ but we fit the model.

$Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + v_i$ and thus commit the specification error of including an unnecessary variable in the model.

(i) Show that the OLS estimators of the parameters of the 'incorrect' model are all unbiased and consistent.

(ii) Show that $\dfrac{\text{var}(\hat{\alpha}_1)}{\text{var}(\hat{\beta}_1)} = \dfrac{1}{1 - r^2_{X_1 X_2}}$ and also show that $\text{var}(\hat{\alpha}_1) \ge \text{var}(\hat{\beta}_1)$.

16. Suppose the true model is $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + u_i$ but you estimate $Y_i = \alpha_0 + \alpha_1 X_i + v_i$.

If you have observations of Y at $X = -3, -2, -1, 0, 1, 2, 3$ and estimate the incorrect model, what bias will result in these estimates ?

17. To see if the variable X_i^2 belongs in the model $Y_i = \beta_0 + \beta_1 X_i + u_i$, Ramsey's RESET test would estimate the linear model, obtaining the estimated Y_i values from this model [i.e., $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$] and then estimate the model $Y_i = \alpha_0 + \alpha_1 X_i + \alpha_2 \hat{Y}_i^2 + v_i$ and test the significance of α_2. Prove that if $\hat{\alpha}_2$ turns out to be statistically significant in the preceding (RESET) equation, it is the same thing as estimating the following model directly : $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$

[**Hint** : Substitute for $\hat{Y}_i$ in the RESET equation.]

18. Consider the model $Y_i = \beta_0 + \beta_1 X_{1i} + u_i$ .... (1)

To find out whether this model is mis-specified because it omits the variable X_2 only. The Lagrange multiplier (LM) test, however, requires you to regress the residuals from model (1) on both X_1 and X_2 and a constant. Why is your procedure likely to be inappropriate ?

19. A researcher tries two specifications of a regression equation : $Y_i = \alpha + \beta X_i + u_i$ and $Y_i = \alpha' + \beta' X_i + \gamma' z_i + u_i'$

Explain under what circumstances the following will be done :

(i) $\hat{\beta} = \hat{\beta}'$

(ii) If $\hat{u}_i$ and $\hat{u}_i'$ are the estimated residuals from the two equations $\Sigma \hat{u}_i^2 \ge \Sigma \hat{u}_i'^2$.

(iii) $\hat{\beta}$ is statistically significant (at the 5% level) but $\hat{\beta}'$ is not.

(iv) $\hat{\beta}'$ is statistically significant (at the 5% level) but $\hat{\beta}$ is not.

20. The model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$ was estimated by ordinary least squares from 26 observations. The results were

$\hat{Y}_i = 2 + 3.5X_{1i} - 0.7X_{2i} + 2.0X_{3i}$

't' ratios : (1.9) (2.2) (1.5) $R^2 = 0.982$

The same model was estimated with the restriction $\beta_1 = \beta_2$. Estimates were

$\hat{Y}_i = 1.5 + 3(X_{1i} + X_{2i}) - 0.6X_{3i}$

't' ratios : (2.7) (2.4) $R^2 = 0.876$

(i) Test the significance of the restriction $\beta_1 = \beta_2$. State the assumptions under which the test is valid.

(ii) Suppose X_{2i} were dropped from the equation. Would the R^2 rise or fall ?

(iii) Would the R^2 rise or fall if X_{1i} were dropped ?

21. The following table shows the values of expenditure on clothing (Y), total expenditure (X_1) and the price of clothing (X_2)

	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969
X_2	16	13	10	7	7	5	4	3	3.5	2
X_1	15	20	30	42	50	54	65	72	85	90
Y	3.5	4.3	5	6	7	9	8	10	12	14

(i) Estimate the model : $Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t$

(ii) Estimate the model : $Y_t = \alpha_0 + \alpha_1 X_{1t} + v_t$

(iii) If $Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t$ is the true model, then examine the consequences on the regression parameters when X_2 is omitted from the model.

22. The following results were obtained from a sample of size 12.

$\Sigma Y_i = 753,\ \ \Sigma Y_i^2 = 48{,}139,\ \ \Sigma X_{1i} Y_i = 40830$

$\Sigma X_{1i} = 643,\ \ \Sigma X_{1i}^2 = 34843,\ \ \Sigma X_{2i} Y_i = 6{,}796$

$\Sigma X_{2i} = 106,\ \ \Sigma X_{2i} = 976,\ \ \Sigma X_{1i} X_{2i} = 5{,}779$

(i) Estimate the model : $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$

(ii) Estimate the model : $Y_i = \alpha_0 + \alpha_1 X_{1i} + v_i$

(iii) Examine the impact on the regression parameters when X_2 is omitted from the true model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$.

APPENDIX

STATISTICAL TABLES

TABLE I

ORDINATES AND AREA OF THE DISTRIBUTION OF STANDARD NORMAL VARIABLE*

τ	ϕ(τ)	Φ(τ)	τ	ϕ(τ)	Φ(τ)	τ	ϕ(τ)	Φ(τ)
.00	.3989423	.5000000	.40	.3682701	.6554217	.81	.2873689	.7910299
.01	.3989223	.5039894	.41	.3667817	.6590970	.82	.2850364	.7938919
.02	.3988625	.5079783	.42	.3652627	.6627573	.83	.2826945	.7967306
.03	.3987628	.5119665	.43	.3637136	.6664022	.84	.2803438	7995458
.04	.3986233	.5159534	.44	.3621349	.6700314	.85	.2779849	.8023375
.05	.3984439	.5199388	.45	.3605270	.6736448	.86	.2756182	.8051055
.06	.3982248	.5239222	.46	.3588903	.6772419	.87	.2732444	.8078498
.07	.3979661	.5279032	.47	.3572253	.6808225	.88	.2708640	.8105703
.08	.3976677	.5318814	.48	.3555325	.6843863	.89	.2684774	.8132671
.09	.3973298	.5358564	.49	.3538124	.6879331	.90	.2660852	.8159399
.10	.3969525	.5398278	.50	.3520653	.6914625	.91	.2636880	.8185887
.11	.2965360	.5437953	.51	.3502919	.6949743	.92	.2612863	.8212136
.12	.3960802	.5477584	.52	.3484925	.6984682	.93	.2588805	.8238145
.13	.3955854	.5517168	.53	.3466677	.7019440	.94	.2564713	.8263912
.14	.3950517	.5556700	.54	.3448180	.7054015	.95	.2540591	.8289439
			.55	.3429439	.7088403	.96	.2516443	.8314724
.15	.3944793	.5596177	.56	.3410458	.7122603	.97	.2492277	.8339768
.16	.3938684	.5635595	.57	.3391243	.7156612	.98	.2468095	.8364569
.17	.3932190	.5674949	.58	.3371799	.7190427	.99	.2443904	.8389129
.18	.3925315	.5714237	.59	.3352132	.7224047	1.00	.2419707	.8413447
.19	.3918060	.5753454	.60	.3332246	.7257469	1.01	.2395511	.8437524
.20	.3910427	.5792597	.61	.3312147	.7290691	1.02	.2371320	.8461358
.21	.3902419	.5831662	.62	.3291840	.7323711	1.03	.2347138	.8484950
.22	.3894038	.5870644	.63	.3271330	.7356527	1.04	.2322970	.8508300
.23	.3885286	.5909541	.64	.3250623	.7389137	1.05	.2298821	.8531409
.24	.3876166	.5948349	.65	.3229724	.7421539	1.06	.2274696	.8554277
.25	.3866681	.5987063	.66	.3208638	.7453731	1.07	.2250599	.8576903
.26	.3856834	.6025681	.67	.3187371	.7485711	1.08	.2226535	.8599289
.27	.3846627	.6064199	.68	.3165929	.7517478	1.09	.2202508	.8621434
.28	.3836063	.6102612	.69	.3144317	.7549029	1.10	.2178522	.8643339
.29	.3825146	.6140919	.70	.3122539	.7580363	1.11	.2154582	.8665005
.30	.3813878	.6179114	.71	.3100603	.7611479	1.12	.2130691	.8686431
.31	.3802264	.6217195	.72	.3078513	.7642375	1.13	.2106856	.8707619
.32	.3790305	.6255158	.73	.3056274	.7673049	1.14	.2083078	.8728568
.33	.3778007	.6293000	.74	.3033893	.7703500	1.15	.2059363	.8749281
.34	.3765372	.6330717	.75	.3011374	.7733726	1.16	.2035714	.8769756
.35	,3752403	.6368307	.76	.2988724	.7763727	1.17	.2012135	.8789995
.36	.3739106	.6405764	.77	.2965948	.7793501	1.18	.1988631	.8809999
.37	.3725483	.6443088	.78	.2943050	.7823046	1.19	.1965205	.8829768
.38	.3711539	.6480273	.79	.2920038	.7852361	1.20	.1941861	.8849303
.39	.3697277	.6517317	.80	.2896916	.7881446	1.21	.1918602	.8868606

TABLE 1 (Contd.)

τ	φ(τ)	Φ(τ)	τ	φ(τ)	Φ(τ)	τ	φ(τ)	Φ(τ)
1.22	.1895432	.8887676	1.71	.0924591	.9563671	2.20	.0354746	.9860966
1.23	.1872354	.8906514	1.72	.0908870	.9572838	2.21	.0347009	.9864474
1.24	.1849373	.8925123	1.73	.0893326	.9581849	2.22	.0339408	.9867906
1.25	.1826491	.8943502	1.74	.0877961	.9590705	2.23	.0331939	.987126
1.26	.1803712	.8961653	1.75	.0862773	.9599408	2.24	.0324603	.9874545
1.27	.1781038	.8979577	1.76	.0847764	.9607961	2.25	.0317397	.9877755
1.28	.1758474	.8997274	1.77	.0832932	.9616364	2.26	.0310319	.9880894
1.29	.1736022	.9014747	1.78	.0818278	.9624620	2.27	.0303370	.9883962
1.30	.1713686	.9031995	1.79	.0803801	.9632730	2.28	.0296546	.9886962
1.31	.1691468	.9049021	1.80	.0789502	.9640697	2.29	.0289847	.9889893
1.32	.1669370	.9065825	1.81	.0775379	.9648521	2.30	.0283270	.9892759
1.33	.1647397	.9082409	1.82	.0761433	.9656205	2.31	.0276816	.9895559
1.34	.1625551	.9098773	1.83	.0747663	.9663750	2.32	.0270481	.9898296
1.35	.1603833	.9114920	1.84	.0734068	.9671159	2.33	.0264265	.9900969
1.36	.1582248	.9130850	1.85	.0720649	.9678432	2.34	.0258166	.9903581
1.37	.1560797	.9146565	1.86	.0707404	.9685572	2.35	.0252182	.9906133
1.38	.1539483	.9162067	1.87	.0694333	.9692581	2.36	.0246313	.9908625
1.39	.1518308	.9177356	1.88	.0681436	.9699460	2.37	.0240556	.9911060
1.40	.1497275	.9192433	1.89	.0668711	.9706210	2.38	.0234910	.9913437
1.41	.1476385	.9207302	1.90	.0656158	.9712834	2.39	.0229374	.9915758
1.42	.1455641	.9221962	1.91	.0643777	.9719334	2.40	.0223945	.9918025
1.43	.1435046	.9236415	1.92	.0631566	.9725711	2.41	.0218624	.9920237
1.44	.1414600	.9250663	1.93	.0619524	.9731966	2.42	.0213407	.9922397
1.45	.1394306	.9264707	1.94	.0607652	.9738102	2.43	.0208294	.9924506
1.46	.1374165	.9278550	1.95	.0595947	.9744119	2.44	.0203284	.9926564
1.47	.1354181	.9292191	1.96	.0584409	.9750021	2.45	.0198374	.9928572
1.48	.1334353	.9305634	1.97	.0573038	.9755808	2.46	.0193563	.9930531
1.49	.1314684	.9318879	1.98	.0561831	.9761482	2.47	.0188850	.9932443
1.50	.1295176	.9331928	1.99	.0550789	.9767045	2.48	.0184233	.9934309
1.51	.1275830	.9344783	2.00	.0539910	.9772499	2.49	.0179711	.9936128
1.52	.1256646	.9357445	2.01	.0529192	.9777844	2.50	.0175283	.9937903
1.53	.1237628	.9369916	2.02	.0518636	.9783083	2.51	.0170947	.9939634
1.54	.1218775	.9382198	2.03	.0508239	.9788217	2.52	.0166701	.9941323
1.55	.1200090	.9394292	2.04	.0498001	.9793248	2.53	.0162545	.9942969
1.56	.1181573	.9406201	2.05	.0487920	.9798178	2.54	.0158476	.9944574
1.57	.1163225	.9417924	2.06	.0477996	.9803007	2.55	.0154493	.9946139
1.58	.1145048	.9429466	2.07	.0468226	.9807738	2.56	.0150596	.9947664
1.59	.1127042	.9440826	2.08	.0458611	.9812372	2.57	.0146782	.9949151
1.60	.1109208	.9452007	2.09	.0449148	.9816911	2.58	.0143051	.9950600
1.61	.1091548	.9463011	2.10	.0439836	.9821356	2.59	.0139401	.9952012
1.62	.1074061	.9473839	2.11	.0430674	.9825708	2.60	.0135830	.9953388
1.63	.1056748	.9484493	2.12	.0421661	.9829970	2.61	.0132337	.9954729
1.64	.1039611	.9494974	2.13	.0412795	.9834142	2.62	.0128921	.9956035
1.65	.1022649	.9505285	2.14	.0404076	.9838226	2.63	.0125581	.9957308
1.66	.1005864	.9515428	2.15	.0395500	.9842224	2.64	.0122315	.9958547
1.67	.0989255	.9525403	2.16	.0387069	.9846137	2.65	.0119122	.9959754
1.68	.0972823	.9535213	2.17	.0378779	.9849966	2.66	.0116001	.9960930
1.69	.0956568	.9544860	2.18	.0370629	.9853713	2.67	.0112951	.9962074
1.70	.0940491	.9554345	2.19	.0362619	.9857379	2.68	.0109969	.9963189
						2.69	.0107056	.9964274

TABLE 1 (Contd.)

τ	φ(τ)	Φ(τ)	τ	φ(τ)	Φ(τ)	τ	φ(τ)	Φ(τ)
2.70	.0104209	.9965330	3.01	.0043007	.9986938	3.30	.0017226	.9995166
2.71	.0101428	.9966358	3.02	.0041729	.9987361	3.31	.0016666	.9995335
2.72	.0098712	.9967359	3.03	.0040486	.9987772	3.32	.0016122	.9995499
2.73	.0096058	.9968333	3.04	.0039276	.9988171	3.33	.0015595	.9995658
2.74	.0093466	.9969280	3.05	.0038098	.9988558	3.34	.0015084	.9995811
2.75	.0090936	.9970202	3.06	.0036951	.9988933	3.35	.0014587	.9995959
2.76	.0088465	.9971099	3.07	.0035836	.9989297	3.36	.0014106	.9996103
2.77	.0086052	.9971972	3.08	.0034751	.9989650	3.37	.0013639	.9996242
2.78	.0083697	.9972821	3.09	.0033695	.9989992	3.38	.0013187	.9996376
2.79	.0081398	.9973646	3.10	.0032668	.9990324	3.39	.0012748	.9996505
2.80	.0079155	.9974449	3.11	.0031669	.9990646	3.40	.0012322	.9996631
2.81	.0076965	.9975229	3.12	.0030698	.9990957	3.41	.0011910	.9996752
2.82	.0074829	.9975988	3.13	.0029754	.9992360	3.42	.0011510	.9996869
2.83	.0072744	.9976726	3.14	.0028835	.9991553	3.43	.0011122	.9996982
2.84	.0070711	.9977443	3.15	.0027943	.9991836	3.44	.0010747	.9997091
2.85	.0068728	.9978140	3.16	.0027075	.9992112	3.45	.0010383	.9997197
2.86	.0066793	.9978818	3.17	.0026231	.9992378	3.46	.0010030	.9997299
2.87	.0064907	.9979476	3.18	.0025412	.9992636	3.47	.0009689	.9997398
2.88	.0063067	.9980116	3.19	.0024615	.9992886	3.48	.0009358	.9997493
2.89	.0061274	.9980738	3.20	.0023841	.9993129	3.49	.0009037	.9997585
2.90	.0059525	.9981342	3.21	.0023089	.9993363	3.50	.0008727	.9997674
2.91	.0057821	.9981929	3.22	.0022358	.9993590	3.51	.0008426	.9997759
2.92	.0056160	.9982498	3.23	.0021649	.9993810	3.52	.0008135	.9997842
2.93	.0054541	.9983052	3.24	.0020960	.9994024	3.53	.0007853	.9997922
2.94	.0052963	.9983589	3.25	.0020290	.9994230	3.54	.0007581	.9997999
2.95	.0051426	.9984111	3.26	.0019641	.9994429	3.55	.0007317	.9998146
2.96	.0049929	.9984618	3.27	.0019010	.9994623	3.56	.0007074	.9998146
2.97	.0048470	.9985110	3.28	.0018397	.9994810	3.57	.0006835	.9998146
2.98	.0047050	.9985588	3.29	.0017803	.9994991	3.58	.0006595	.9998282
2.99	.0045666	.9986051				3.59	.0006343	.9998347
3.00	.0044318	.9986501				3.60	.0006119	.9998409

* Abridged from Table 1 of *Biometrika Tables for Statisticians*, vol. I, with the kind permission of the Biometrika Trustees.

TABLE II

DISTRIBUTION OF STANDARD NORMAL VARIABLE

Values of τ_α

α	0.05	0.025	0.01	0.005
τ_α	1.645	1.960	2.326	2.576

TABLE III
χ^2-DISTRIBUTION*
VALUES OF $\chi^2_{\alpha,\nu}$

ν \ α	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.070	12.832	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.688	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.558
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672
40	20.706	22.164	24.433	26.509	55.759	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	67.505	71.420	76.154	79.490
60	35.535	37.485	40.482	43.188	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	124.342	129.561	135.807	140.169

For larger values of v, the quantity $\sqrt{2\chi^2}-\sqrt{2v-1}$ may be used as a standard normal variable.

*Abridged from Table 8 of *Biometrika Tables for Statisticians*, vol. I, with the kind permission of the Biometrika Trustees.

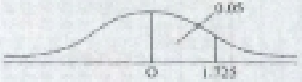
TABLE IV
t-DISTRIBUTION*
VALUES OF $t_{\alpha,\nu}$

Example

$\Pr(t > 2.086) = 0.025$

$\Pr(t > 1.725) = 0.05$ for df = 20

$\Pr(|t| > 1.725) = 0.10$



Pr(α) / df(ν)	0.25 0.50	0.10 0.20	0.05 0.10	0.025 0.05	0.01 0.02	0.005 0.010	0.001 0.002
1	1.000	3.078	6.314	12.706	31.821	63.657	318.31
2	0.816	1.886	2.920	4.303	6.965	9.925	22.327
3	0.765	1.638	2.353	3.182	4.541	5.841	10.214
4	0.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	0.718	1.440	1.943	2.447	3.143	3.707	5.208
7	0.711	1.415	1.895	2.365	2.998	3.499	4.785
8	0.706	1.397	1.860	2.306	2.896	3.355	4.501
9	0.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.700	1.372	1.812	2.228	2.764	3.169	4.144
11	0.697	1.363	1.796	2.201	2.718	3.106	4.025
12	0.695	1.356	1.782	2.179	2.681	3.055	3.930
13	0.694	1.350	1.771	2.160	2.650	3.012	3.852
14	0.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	0.690	1.337	1.746	2.120	2.583	2.921	3.686
17	0.689	1.333	1.740	2.110	2.567	2.898	3.646
18	0.688	1.330	1.734	2.101	2.552	2.878	3.610
19	0.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	0.686	1.323	1.721	2.080	2.518	2.831	3.527
22	0.686	1.321	1.717	2.074	2.508	2.819	3.505
23	0.685	1.319	1.714	2.069	2.500	2.807	3.485
24	0.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.684	1.316	1.708	2.060	2.485	2.787	3.450
26	0.684	1.315	1.706	2.056	2.479	2.779	3.435
27	0.684	1.314	1.703	2.052	2.473	2.771	3.421
28	0.683	1.313	1.701	2.048	2.467	2.763	3.408
29	0.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.683	1.310	1.697	2.042	2.457	2.750	3.385
40	0.681	1.303	1.684	2.021	2.423	2.704	3.307
60	0.679	1.296	1.671	2.000	2.390	2.660	3.232
120	0.677	1.289	1.658	1.980	2.358	2.617	3.160
∞	0.674	1.282	1.645	1.960	2.326	2.576	3.090

Note : The smaller probability shown at the head of each column is the area in one tail, the larger probability is the area in both tails.

Source : From E. S. Pearson and H. O. Hartley, eds., *Biometrika Tables for Statisticians*, vol. 1, 3rd ed., table 12. Cambridge University Press, New York. 1996. Reproduced by permission of the editors and trustees of *Biometrika*.

TABLE V
F-DISTRIBUTION*
Values of $F_{0.05;\ v_1, v_2}$

$v_2 \backslash v_1$	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.2	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

For other values of v_1 and v_2, one may use linear interpolation, taking $1/v_1$ and $1/v_2$ as the independent variables.

TABLE V (Contd.)
Values of $F_{0.01;\ v_1, v_2}$

$v_2 \backslash v_1$	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999.5	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

For other values of v_1 and v_2, one may use linear interpolation, taking $1/v_1$ and $1/v_2$ as the independent variables.

*Abridged from Table 18 of Biometrika Tables for Statisticians, vol. I, with the kind permission of the Biometrika Trustees.

TABLE VI
THE DURBIN-WATSON d-STATISTIC
SIGNIFICANCE POINTS OF d_L AND d_U : 5%

	$k' = 1$		$k' = 2$		$k' = 3$		$k' = 4$		$k' = 5$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
6	0.61	1.40	—	—	—	—	—	—	—	—
7	0.70	1.36	0.46	1.89	—	—	—	—	—	—
8	0.76	1.33	0.55	1.77	0.36	2.28	—	—	—	—
9	0.82	1.32	0.62	1.69	0.45	2.12	0.29	2.58	—	—
10	0.87	1.32	0.69	1.64	0.52	2.01	0.37	2.41	0.24	2.82
11	0.92	1.32	0.65	1.60	0.59	1.92	0.44	2.28	0.31	2.64
12	0.97	1.33	0.81	1.57	0.65	1.86	0.51	2.17	0.37	2.50
13	1.01	1.34	0.86	1.56	0.71	1.81	0.57	2.09	0.44	2.39
14	1.04	1.35	0.90	1.55	0.76	1.77	0.63	2.03	0.50	2.29
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.88	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1,12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.60	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.41	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78
150	1.72	1.74	1.70	1.76	1.69	1.77	1.67	1.78	1.66	1.80
200	1.75	1.77	1.74	1.78	1.73	1.79	1.72	1.81	1.71	1.82

Note : k' = Number of explanatory variables excluding the constant.